Psychometrika

CONTENTS

VOLUME TWENTY-TWO DECEMBER 1957 NUMBER 4

PSYCHOMETRIKA—VOL. 22, NO. 4
DECEMBER, 1957

NEW PROBLEMS FOR OLD SOLUTIONS*

HUBERT E. BROGDEN
PERSONNEL RESEARCH BRANCH
THE ADJUTANT GENERAL'S OFFICE†

Some of the methodologies that have become standard tools in psychometrics suffer from neglect. They are taken too much for granted and are not given the attention that seems appropriate to the important role they play in research advances. I propose to make some suggestions which may, in a modest way, assist in alleviating this difficulty. In a very general way I would like to suggest that effort expended in examining a variety of restatements of a methodological problem may lead to new methodologies of real value.

In a sense then, I am suggesting that we look for new problems in areas where old solutions are available, or possibly, for problem restatements where the application of an old solution may have become too automatic and too uncritical.

Since a discussion of all of this in general terms would tend toward triteness, I plan to proceed in part through the use of examples with the thought in mind that a number of such examples (bearing upon similar problems) may be of some additional value in suggesting a generalized approach to certain classes of problems. For simplicity, I will avoid problems associated with sampling error and limit the discussion throughout the paper to cases involving very large samples.

An article by Guilford and Michael entitled "Approaches to Univocal Factor Scores" illustrates the kind of problem restatement that I have in mind. A solution to the problem of estimating factor scores, providing estimates that are best in the least square sense, has been available for some time. However, it has been frequently observed that the least squares estimates of orthogonal factors tend to intercorrelate substantially. Having in mind, possibly, this apparent defect of scores estimated through the least squares method, Guilford and Michael suggest as an alternative approach scoring or weighting procedures designed to yield a univocal factor score—a score having variance in only one common factor, its remaining variance being

*Presidential Address to the Psychometric Society, September 4, 1957.

†The opinions expressed are those of the author and do not reflect official Department of the Army policy.

attributable to errors of measurement plus possible specific variance. This restatement emphasizes the reduction of bias or contamination and relegates accuracy of measurement to a secondary role.

While I do not wish to consider this problem in greater detail or attempt to describe the solution, I do have a further point regarding the justification of the problem restatement, which is pertinent here and will be pertinent in the examples to be discussed later. In considering possible methods for estimating factor scores, an early question might well be—how will the factor scores be used, or what kinds of conclusions will be drawn and what kinds of decisions will be made as a result of their use? It seems desirable that the statement of the problem should be phrased so as to maximize the likelihood that such conclusions or decisions will be correct. This may often lead to a number of alternate problem statements since the same problem statement may not permit correct conclusions in research studies with various objectives.

Let us consider, for a moment, the factor score problem in relation to these questions. I realize that many investigators seeking estimates of factor scores may have purposes in mind which are consistent with the use of the least squares estimates. They may wish, for example, to report estimates of the factor scores to the individual subjects taking a set of tests. To illustrate the role of anticipated conclusions or decisions in the statement of the problem, let us consider a class of studies in which the investigator seeks to extend his knowledge of a set of factors by relating the factor score estimates to a set of new variables. This will permit us to relate the conclusions resulting from such a study to the statement of the problem.

An investigator with such a purpose in mind will draw no conclusions from the individual scores and has no proper direct interest in such scores. He will draw conclusions only after correlating the factor scores with the additional set of variables under investigation. His conclusions will relate specifically to the correlations between the factor estimates and this additional set of variables. A problem restatement such as that of Guilford and Michael could well have been tied directly to this intended use of the factor estimates. If the factor estimates are collinear with the factor, the conclusions can be proper and valid; if they are not collinear, the correlations will be biased (as estimates of factor loadings) and the conclusions will be directly and adversely affected.

A consideration of the problem of estimating a composite criterion score from unreliable components will provide a further opportunity to illustrate the kind of problem restatement I have in mind. Suppose that a set of criterion components are available which are deficient only in that error of measurement is present. The problem, in general, is the selection of weights for the components that will yield the best estimate of the true composite.

A least squares solution to this problem is possible and has often been proposed as best. If, however, we reconsider the statement of the problem and examine particularly the nature of the conclusions to be drawn when the estimate of the criterion composite is used for validation research, it can be shown that the least squares solution is irrelevant and gives weights negatively related to those provided by the proper solution.

Assuming that the criterion composite is to be used for validation studies, the validity coefficients or the partial regression weights of the predictors are of central importance, since these are basic to the conclusions deriving from the validation study. The following restatement of the problem is suggested after considering the intended use of the criterion: what set of weights for the fallible criterion components will insure that, in a later validation study, the validity coefficients (or the partial regression weights) obtained against the estimated composite will be the same as those that would be obtained if the true composite were available?

R. H. Gaylord and I have considered this problem and a solution has been achieved. An interesting aspect of the solution is the relationship between the magnitude of the weights and the reliability of the components—the less reliable the component the larger the weight, when the components are in standard score form. To understand this point note that—when the components have unit variance—if more error variance is present less true score variance will remain. Hence, a heavier weight is needed if the true score variance of an unreliable component is to be proportionately represented in the over-all criterion composite.

With a least squares solution the opposite is true: the greater the error in the criterion component the lower the least squares weight. The least squares methods yield a composite that has maximum correlation with the true composite but which is biased for the purpose of validation research.

The approaches to this problem and to the problem of factor score estimation have much in common. In each instance the least squares solution was, perhaps, the more obvious one. The problem restatement could in each case be derived from an examination of the way in which the solution would affect the conclusions of research studies in which it was applied. There is a further point of similarity. The criterion estimation problem might have been stated: what estimated criterion composite will be collinear with the true composite?

A restatement of the problem of item difficulty distribution may throw still further light on the general point I have in mind. A number of investigators have in various ways studied the relation of tests to underlying ability and have shown for various conditions the characteristics of items and item difficulty distributions that will yield the most efficient measurement of underlying ability. While efficiency of measurement has been defined in a number of ways, all of the definitions resemble the least squares definition in that all are concerned with maximizing some index of the degree of relationship between the test and underlying ability. Ferguson and others have discussed the problem associated with difficulty factors and have stressed the way in which correlations among tests may be distorted as a function of similarity or dissimilarity in the difficulty distribution of the items

Since the phenomenon of difficulty bias in correlations is basic to the problem restatement I wish to propose, further explanation of this bias seems desirable at this point. It is well known that if the p-values of two dichotomous variables are similar, the phi coefficient will tend to be low and this will hold although the tetrachoric correlations between all pairs of items are equal. Now, with test scores the same phenomenon is evident, particularly if the tests are homogeneous in difficulty. Two tests, each homogeneous in difficulty, will correlate more highly if the difficulty level of the two tests is approximately the same than if the difficulty levels are divergent. Many test types proposed as efficient measures of underlying ability in the least squares sense have items homogeneous in difficulty. Hence, the correlations among such tests and between such tests and other variables are also subject to difficulty bias—probably more so than with tests having a greater spread of item difficulty.

I do not mean to disparage tests designed as efficient measures of underlying ability or to imply that the statement of the problem leading to this solution is defective. I merely wish to suggest that an alternate problem statement is possible and, for certain purposes, may be more desirable.

Where a contribution to subject matter knowledge is the object of an investigation, it seems proper that elimination of bias is all-important and reduction of error of measurement is of secondary importance particularly since, with knowledge of error of measurement, methods of estimation are often available and appropriate which will make allowance for the attenuating effect of error.

Thus the problem, as restated in a general manner, might be to determine the optimum difficulty distribution for a test to be used for investigating the relation between the ability it measures and other variables. What we want is a test which will yield the same pattern of correlations with other variables as would underlying ability, regardless of the nature of such other variables. If the correlations between the test and other variables are the same, after correcting for the effect of error, as the correlations between underlying ability and such variables, the conclusions or decisions reached will be the same.

I cannot offer, with proof, an exact statement of the over-all problem and an accompanying solution. While the foregoing discussion is possibly sufficient in view of the theme and scope of this paper, a further problem statement and a possible solution may be of some interest. The justification of these further developments must remain largely intuitive.

A brief indication of the major assumptions and limiting conditions may be helpful before continuing with the new problem statement. Obviously, in this brief discussion it is not feasible to state these in full. Underlying ability is defined as a perfect normally distributed measure of the ability common to the dichotomous items. The problem is limited to tests in which all items have the same biserial correlation with underlying ability. Ability and error are assumed to be the only determiners of the item responses.

If I rephrase the problem statement and ask: what item difficulty distribution will yield a test score such that the bivariate frequency surface of the test score and underlying ability is normal, the statement then appears to be more precise and seems to be a more feasible starting point in a mathematical development leading to a demonstrated solution.

This more precise restatement is, I believe, logically equivalent to the prior and more general restatement of the problem. From this second restatement it follows that the test is a simple linear function of underlying ability and error, and that the test is described thus through its entire range—given only the product moment correlation between the test and ability. It also follows, then, that the correlation between ability and any other variable can be estimated, given the correlation between this variable and the test and, of course, the correlation between ability and the test. A linear model equivalent to that used in factor analysis is applicable and the estimate is the product of the above two correlations. It is emphasized that this model is believed to hold regardless of difficulty biases that may be present in the other variable. Hence, when the bivariate surface and ability is normal, it seems reasonable that the test can be used in place of ability and, with correction for error of measurement, the conclusions reached through the use of the test are the same as those that would be reached had underlying ability been available.

The item difficulty distribution that I have in mind as a possible solution to this problem is perfectly rectilinear, with the item difficulty index expressed as baseline values of a normal curve. To achieve this distribution, items would be selected with difficulties of 0, +.1, −.1, +.2, −.2, +.3, −.3, etc. In theory, such a distribution would place items at equal difficulty intervals ranging from plus infinity to minus infinity.

Now, the general point of this paper had to do with the value of examining alternate statements of the problem, and I believe these three examples illustrate this general point. I have been developing as a second general point the need for examining the conclusions or decisions to be made when the methodologies under consideration are to be applied, and the desirability of reasoning from such conclusions to a justification of the methodology. This point has been stressed sufficiently and discussed in relation to each of the three examples.

The distinction between the kind of problem statement that leads to a least squares solution (or something akin to such a solution) and the alternate problem statements that we have considered deserves some extra comment.

We have noted that if, in each case, the scores were to be used for practical estimation of an individual's standing, the least squares solution would likely be satisfactory.

If interest did not center on direct use of the scores and if the scores were to be used as a means of arriving at further conclusions through additional research, alternate problem statements pointing toward reduction of bias have been suggested as more pertinent and more acceptable.

The added suggestion I wish to make at this point deviates from the central theme of the paper and relates to the above similarities in the three examples. I am suggesting merely that the above-noted distinction may extend beyond these three examples. In additional problems where the least squares solution has been accepted as best, a close examination of the problem in relation to the decisions to be made when the resulting method is used may again suggest an alternate problem statement. In other words, the particular distinctions between the least squares problem statement and the reduction of bias problem statement may have more general value.

The latter portion of this paper will be directed toward possible sources of confusion between different classes of problems or problem statements rather than toward restatements of problems as such.

A very general distinction in the methodologies widely used in psychology is relevant in several ways to the present discussion, although little of what I have to say is really new or different. I am speaking of the distinction between a correlational approach and approaches primarily based on controlled experimentation. I would like to discuss these two very general approaches in relation to classes of practical decisions properly stemming from empirical evidence. I am choosing cases involving practical decisions so that certain points can be made most clearly, not because the points I wish to make are necessarily limited to cases involving practical decisions.

If the practical decision is a choice between administering or not administering a given treatment, it is well recognized that a controlled experiment is properly used to demonstrate the effect of the treatment. Knowledge of the effect of the treatment then becomes a major factor in deciding whether or not the treatment will be used. I have no real comment here. I believe that few would hold that a correlational study—without experimental controls—is proper backing for such a decision.

While the class of decision for which correlational evidence is appropriate is fairly well recognized in practice, it is somewhat more difficult to find a clear statement enjoying widespread agreement in discussion of scientific method. I should like to suggest at least one type of practical decision where a correlational design is clearly pertinent. I mean, specifically, a decision to use or not to use a test or measure for the identification and hiring of personnel. With regard to this type of decision, I would like to make two points:

(1) a correlation design will show what criterion performance can be expected from persons with a given test score, thus giving information basic to the decision in question,

and

(2) a controlled experiment (showing the relationship between a test and an appropriate criterion with other variables held constant) may suggest but does not demonstrate the value of this independent variable for selection purposes.

My point regarding the correlation design should be clear and acceptable without elaboration. The second point calls for further discussion.

A true controlled experiment is in some ways quite meaningless when a test of an ability or personality trait is the independent variable. A test score cannot be meaningfully manipulated—the individual differences in the test score must be taken as they come or created by selection of cases. Moreover, experimental controls are difficult to accomplish. Such controls must again be achieved by selection of cases. Most important, however, it is difficult to define the "other variables" that are to be held constant in a controlled experiment. Consider, for example, the consequences of holding constant an alternate form of the test used as the independent variable, or the consequences of holding constant a number of tests so chosen that the common-factor variance of the independent variables will be reduced to zero.

If we disregarded the problems I have just raised and assume that a test has been found to predict a criterion, and that all other variables were held constant through selection of cases, an additional difficulty still arises. With selection of cases the sample in which this relationship is demonstrated can no longer resemble the sample in which the application must take place, and the relationship discovered in the validation sample cannot be applied with confidence. To further clarify this point, consider the kind of selection procedure that *is* supported by the evidence of a controlled experiment. The two steps of the procedure are: (1) selection of applicants to duplicate the effect of the operations used in the validation sample to hold all other variables constant, and (2) within the remaining applicants, selection of those with high scores on the test under investigation. Needless to say, this two-step procedure is not appropriate to the practical problem.

Although, as I had suggested earlier, the thoughts expressed with regard to these two designs are not new, I hope that the examination of the designs in relation to the decisions to which they are pertinent may have provided some new insight into the distinction between these problems.

A second general distinction between classes of problems arises in connection with scaling. Consider, as one problem, the search for units of measurement that have the properties of a true scale. Many authors have struggled with this problem as it relates to the general methodology of science and most agree to a number of desirable features of so-called true scales. Such scales are fairly common in the physical sciences. In psychology, they are sought after but rarely achieved.

A second class of scaling problem is, I believe, distinctly different from the problems associated with true scales. If the scaling problem is pointed toward practical application and tied to decision theory, we are then seeking units such that the scale has the properties necessary to permit the decision in question. Thus, particularly in connection with criterion problems for personnel selection research, the notion of a common metric and the notion of equal units of scaling deviates from the philosophy behind true scales and takes on a very different meaning. It may well be that a dollar unit is a true common metric and a true scale unit in all of the senses relevant to decisions properly made as a result of selection research.

Yet such a scale, while highly relevant to this kind of decision, has no apparent relevance to the scaling problems mentioned earlier. If we construct a performance test to measure—job sample wise—proficiency in a particular job element, we should not seek a scale that discriminates equally well at all levels of difficulty and conforms to the desirable attributes of a true scale. We seek a count of behaviors that is representative of the actual behavior in this particular aspect of the job. Such a scale will discriminate only at the appropriate level of difficulty whether the behaviors required of the job incumbent are all very easy or all very difficult. In other words, we seek a count of behaviors, and we seek them at a difficulty level such that they can be properly evaluated as representing profits or losses to a decision maker—assuming that the purpose of the decision maker is to maximize profits.

I suspect that this difference in the purposes of scaling—as seen in a practical decision problem on the one hand and in the development of a general body of scientific knowledge on the other—can be differentiated further. I suspect, also, that many investigators have not distinguished between these two types of scaling problems and that the scales developed may have been less adequate or less suitable as a result.

In summary, let me point again to several of the major points of this paper. Let me repeat and emphasize my belief that, in developing a methodology, we must closely examine the decisions to be made or the conclusions to be drawn when the methodology is applied. The methodologies should be molded so that a correct decision can follow an application of the methodology, and the chain of reasoning, in my opinion, best proceeds from a definition of the decisions to be made to a justification of the methodology.

The three examples involving a distinction between the least squares solution and solutions offering measures free of bias have suggested a second point. I believe that this distinction can be usefully applied in other contexts and that some added insight will obtain.

Finally, I hope that the foregoing has clarified my most general thesis and that I have given some support to the notion that effort expended in seeking new problem statements can be profitable.

Manuscript received 9/4/57

PSYCHOMETRIKA—VOL. 22, NO. 4
DECEMBER, 1957

OPTIMAL TEST LENGTH FOR MULTIPLE PREDICTION: THE GENERAL CASE*

PAUL HORST AND CHARLOTTE MACEWAN
UNIVERSITY OF WASHINGTON

The concepts of differential prediction and multiple absolute prediction were developed in earlier papers [2, 3]. Methods for determining optimal distribution of testing time for each type of prediction are available [4, 5] and are appropriate for use provided that no altered time allotment approaches zero. In this article the methods developed in [4, 5] are extended to include cases where the altered time allotment for one or more tests may approach zero. The procedures developed are illustrated by numerical examples, after which the mathematical rationales are provided.

In previous publications [2, 3, 4, 5] the problem of maximum validity in predicting multiple criteria was approached in two different ways. In [2] and [3], for predicting criteria differentially and for multiple absolute prediction, respectively, techniques were developed for selecting from a large number of potential predictors that subset, of specified size, which yields the highest over-all validity as measured by the respective indices of prediction efficiency, ϕ and λ. A more general approach was used in [4] and [5], in which a procedure previously presented for the case of a single criterion [1] was extended to the cases of differential prediction and multiple absolute prediction, respectively. Here, techniques were presented whereby, starting with a given battery of predictors for differential prediction or for multiple absolute prediction, one could determine altered administration time allotments, for any specified over-all testing time, for which the index of prediction efficiency (ϕ or λ, respectively) would be a maximum.

The techniques developed in [4] and [5] provide methods of solving for optimal test lengths, in terms of time allotments, by series of approximations. Since reciprocals of the altered time allotments are involved, the methods do not hold in the event that any altered testing time becomes zero. In this article a modification of procedure, applicable also in the case in which the new time allotment for any test approaches zero, is presented.

A numerical example for the case of differential prediction, and a summary for the case of multiple absolute prediction follow in the next section. The mathematical basis for the procedures described for the general case is presented in the final section.

*This research was carried out under Contract Nonr-477(08) between the University of Washington and the Office of Naval Research. The authors express their appreciation to Shun Mei Ling for carrying out the computations, and to Elizabeth Cross for typing the manuscript for publication.

Numerical Examples

The General Case for Differential Prediction

The example below demonstrates a modification of the computational procedure presented in [4] such that its applicability is perfectly general. The assumptions stated for the more restricted case [1, 4] also apply in the general case but will not be repeated here.

The data used in this example are those used in [4]. The matrix of test intercorrelations with reliabilities in the diagonal is shown in Table 1. Criterion variables are grade-point averages in each of ten college areas. The matrix of validity coefficients is shown in Table 2.

Over-all testing time for the tests of arbitrary length is 142 minutes. Assume, as was the case for the example in [4], that the total testing time is to be cut in half, that is, to 71 minutes. The problem is to determine time allotments for the various tests such that the resulting index of differential prediction efficiency is maximized. The following method of solution employs a series of approximations differing from that presented in [4]—with the exception of the first iteration, no reciprocals of the altered time allotments are involved.

It will be demonstrated that the results obtained by the original and the modified procedures are, for practical purposes, virtually identical. Since we start with no near-zero test lengths, the somewhat shorter method, as described by steps 1–8f in [4] may be used to obtain the second approximation to optimal test lengths. In brief, by these steps we determine:

1. The α_c' matrix shown in Table 3. This is obtained from Table 2 by subtracting the mean of column i from each element in column i.

2. The elements in the diagonal matrix, Δ, shown in row 2 of Table 4. Each element is the original test length, given in row 1 of Table 4, multiplied by the corresponding unreliability.

3. The first approximation to the altered test lengths. Assume each test length cut in half as shown in row 3 of Table 4.

4–5. The values shown in row 4 of Table 4. Each element in row 2 of Table 4 is divided by the corresponding element in row 3 of Table 4.

6–7. The matrix L_1 . To compute the matrix L_1 , first make up a matrix as follows: Using the R matrix of Table 1, the value of each diagonal element is increased by adding to it the value of the corresponding element in row 4 of Table 4. For example, the first diagonal element of the new matrix is $.920 + .160 = 1.080$. The L_1 matrix is obtained by premultiplying the matrix α_c of Table 3 by the inverse of the augmented R matrix. The procedure for premultiplying a matrix by the inverse of a symmetric matrix is outlined in [6]. The solution is found in two stages, the "forward solution" and the "backward solution," both of which may be seen in [4]. In this report only the

TABLE 1

The R Matrix of Predictor Intercorrelations with Reliabilities

Substituted for Unities in the Diagonal:

$R = r - D_u$

		1	2	3	4	5	6	Σ
1	G-Z 1	.920	.159	.152	.281	.763	.515	2.790
2	G-Z 3	.159	.920	.003	.369	.292	.243	1.986
3	G-Z 7	.152	.003	.920	.200	.142	-.150	1.267
4	ACE-Q	.281	.369	.200	.820	.549	.426	2.645
5	ACE-L	.763	.292	.142	.549	.830	.628	3.204
6	English	.515	.243	-.150	.426	.628	.860	2.522
Σ		2.790	1.986	1.267	2.645	3.204	2.522	14.414

TABLE 2

The r_c Matrix of Validity Coefficients

		1 G-Z 1	2 G-Z 3	3 G-Z 7	4 ACE-Q	5 ACE-L	6 English	Σ
1	Anthropology	.370	.177	.091	.294	.341	.357	1.630
2	Chemistry	.317	.274	.016	.309	.364	.399	1.679
3	Economics	.339	.211	.008	.241	.334	.323	1.456
4	English	.526	.247	-.075	.262	.488	.524	1.972
5	Foreign Lang.	.295	.287	-.156	.200	.232	.426	1.284
6	Geology	.184	.140	.094	.170	.229	.214	1.031
7	History	.379	.169	-.001	.182	.373	.336	1.438
8	Mathematics	.287	.348	-.088	.350	.336	.401	1.634
9	Psychology	.440	.170	.096	.285	.409	.403	1.803
10	Zoology	.336	.216	.031	.318	.345	.351	1.597
Σ		3.473	2.239	.016	2.611	3.451	3.734	15.524
Σ/10		.347	.224	.002	.261	.345	.373	1.552

TABLE 3

The a_c' Matrix: Validity Coefficients Expressed

in Deviation Form for Each Test

	1	2	3	4	5	6
1	.023	-.047	.089	.033	-.004	-.016
2	-.030	.050	.014	.048	.019	.026
3	-.008	-.013	.006	-.020	-.011	-.050
4	.179	.023	-.077	.001	.143	.151
5	-.052	.063	-.158	-.061	-.113	.053
6	-.163	-.084	.092	-.091	-.116	-.159
7	.032	-.055	-.003	-.079	.028	-.037
8	-.060	.124	-.090	.089	-.009	.028
9	.093	-.054	.094	.024	.064	.030
10	-.011	-.008	.029	.057	.000	-.022
Ck	.003	-.001	-.004	.001	.001	.004
Σ	.003	-.001	-.004	.001	.001	.004

TABLE 4

The $1'D_a$ Row Vector of Original Test Lengths, and the $1'D_{b_1}$

Row Vector of First Approximations to Optimal Test

Lengths in Minutes, and $1'\Delta$

		1	2	3	4	5	6	Σ
1	$1'D_a$	25.0	9.0	30.0	23.0	15.0	40.0	142.0
2	$1'\Delta$	2.000	.720	2.400	4.140	2.550	5.600	17.410
3	$1'D_{b_1} = \frac{1}{2}1'D_a$	12.5	4.5	15.0	11.5	7.5	20.0	71.0
4	$1'\Delta D_{b_1}^{-1}$	.160	.160	.160	.360	.340	.280	1.460

backward solution, in Table 5, showing the transpose of the L_1 matrix in the upper left section is reproduced.

8. The second approximation to the altered test lengths. The computational procedure is that used in [4] and is shown in rows *a* through *f* in Table 5.

Row *a* consists of the sums of squares of column elements of the L_1' matrix. For example, the first element in row *a*, .0626, is the sum of squares of the first 10 elements in column 1 of Table 5.

Row *b* is copied from row 2 of Table 4.

Row *c* consists of the products of corresponding elements in the two preceding rows. For example, the first element in row *c*, .1251, is $.0626 \times 2.00$. (In the original computations, six decimals were retained in the elements of row *a*.)

Row *d* consists of the square roots of the corresponding elements in the preceding row. For example, the first element is $\sqrt{.1251} = .3537$. The value of *s*, as seen to the right of this row, is computed as the over-all new testing time, 71 minutes, divided by the sum of elements in row *d*, 1.8823. The quotient is 37.7198.

Row *e* gives a check on the computations for row *d*. Each element in row *c* is divided by the corresponding element in row *d*. Thus, $.1251/.3537 = .3537$.

Row *f* has as elements the second approximations to optimal test lengths. These values are found by multiplying each element in row *d* by the obtained value of *s*. Thus, for the first element, $.3537 \times 37.7198 = 13.3415$. Summed, the values in row *f* should equal 71, the over-all new testing time in minutes.

Since there are no near-zero values in row *f*, normally one would continue in the manner described in [4] to obtain the third approximation to altered test lengths; i.e., in terms of the present report, substitute the values in row *f* of Table 5 for those in row 3 of Table 4, and repeat steps 4–5 through 8*f* to compute the third approximation to optimal test lengths.

Assume, on the contrary, that some test length as given in row *f* of Table 5 were near-zero or zero. Under these conditions, it would be difficult or impossible, in the succeeding iteration to carry out the computations indicated in step 4–5. The modified procedure described below avoids such an impasse. This procedure may be employed with complete generality. The calculations in Table 5 are completed as follows.

Row *g* consists of the square roots of the corresponding elements in the preceding row. Thus, for the first element, $\sqrt{13.3415} = 3.6526$.

Row *h* gives a check on the calculation of row *g*. Each element in row *f* is divided by the corresponding element in row *g*. For example, $13.3415/3.6526 = 3.6526$.

The elements of row *g* will be used subsequently for a number of operations.

TABLE 5

Computation of $(R+\Delta D_{b_1}^{-1})^{-1} a_c = L_1$ (Backward Solution) from (4, p. 58), and

Computations for the Second Approximation to Optimal Test Lengths

L_1' Matrix

	1	2	3	4	5	6	1	2	3	4	5	6	7	8	9	10	Check		$L_1 1$
1	.039	-.051	.074	.044	-.044	-.003	-1										0	.001	.255
2	-.075	.035	.019	.021	.030	.027		-1									0	-.001	.207
3	.008	-.004	-.005	-.005	.020	-.056			-1								0	.003	.098
4	.136	-.002	-.075	-.058	.036	.063				-1							0	.000	.370
5	.031	.085	-.110	-.026	-.166	.101					-1						0	.001	.519
6	-.159	-.040	.101	-.050	.058	-.059						-1					0	.000	.467
7	.022	-.039	-.011	-.079	.087	-.053							-1				0	.002	.291
8	-.068	.102	-.087	.082	-.018	.001								-1			0	.002	.358
9	.074	-.067	.078	.003	.006	.013									-1		0	.002	.241
10	-.005	-.021	.010	.070	-.008	-.033										-1	0	-.001	.147

		1	2	3	4	5	6	Σ	Check	
a	$1'D_{L_1L_1'}$	.0626	.0295	.0477	.0268	.0434	.0256			
b	$1'\Delta$	2.00	.72	2.40	4.14	2.55	5.60	17.41	17.41	
c	$1'D_{L_1L_1'}\cdot\Delta$	.1251	.0214	.1144	.1110	.1108	.1433			
d	$1'(D_{L_1L_1'}\cdot\Delta)^{\frac{1}{2}}$	.3537	.1458	.3382	.3332	.3328	.3786	1.8823		$s = \frac{T_1}{1'(D_{L_1L_1'}\cdot\Delta)^{\frac{1}{2}}1} = \frac{71}{1.8823} = 37.7198$
e	Ck: $1'D_{L_1L_1'}\cdot\Delta(D_{L_1L_1'}\cdot\Delta)^{-\frac{1}{2}}$	.3537	.1458	.3382	.3332	.3329	.3786			
f	$1'D_{b_2} = 1'(D_{L_1L_1'}\cdot\Delta)^{\frac{1}{2}}s$	13.3415	5.4995	12.7568	12.5682	12.5531	14.2807	70.9998	71.0000	
g	$1'D_{b_2}^{\frac{1}{2}}$	3.6526	2.3451	3.5717	3.5452	3.5430	3.7790	20.4366		
h	Ck: $1'D_{b_2}D_{b_2}^{-\frac{1}{2}}$	3.6526	2.3451	3.5716	3.5451	3.5431	3.7790			

9. The matrix shown in Table 6 is calculated next. Each *column* of Table 6 is obtained by multiplying each element in the corresponding *column* of the R matrix shown in Table 1 by the corresponding element in row g. For example, for the first column of Table 6, the first two elements are: .920 × 3.6526 = 3.3604; .159 × 3.6526 = .5808.

TABLE 6

The $RD_{b_2}^{\frac{1}{2}}$ Matrix

	1	2	3	4	5	6	Σ
1	3.3604	.3729	.5429	.9962	2.7033	1.9462	9.9219
2	.5808	2.1575	.0107	1.3082	1.0346	.9183	6.0101
3	.5552	.0070	3.2860	.7090	.5031	.5669	4.4934
4	1.0264	.8643	.7143	2.9071	1.9451	1.6099	9.0681
5	2.7869	.6848	.5072	1.9463	2.9407	2.3732	11.2391
6	1.8811	.5699	-.5358	1.5103	2.2250	3.2499	8.9004
Σ	10.1908	4.6574	4.5253	9.3771	11.3518	9.5306	49.6330
Ck	10.1908	4.6574	4.5253	9.3771	11.3518	9.5306	49.6330

TABLE 7

The $D_{b_2}^{\frac{1}{2}} RD_{b_2}^{\frac{1}{2}}$ Matrix

	1	2	3	4	5	6	Σ	Ck
1	12.274	1.362	1.983	3.639	9.874	7.109	36.241	36.241
2	1.362	5.060	.025	3.068	2.426	2.154	14.095	14.094
3	1.983	.025	11.737	2.532	1.797	-2.025	16.049	16.049
4	3.639	3.068	2.532	10.306	6.896	5.707	32.148	32.148
5	9.874	2.426	1.797	6.896	10.419	8.408	39.820	39.820
6	7.109	2.154	-2.025	5.707	8.408	12.281	33.634	33.635

10. Computed next is the matrix found in Table 7. Each *row* of Table 7 is obtained by multiplying each element in the corresponding *row* of the table computed in step 9 by the corresponding element in row g. For example, elements one and two of row 1 of Table 7 are: 3.3604 × 3.6526 = 12.274; .3729 × 3.6526 = 1.362.

11. Calculate a matrix which shall be designated A_1 . The Δ values found in row 2 of Table 4 are added to the corresponding diagonal elements of the table obtained in step 10, and the resulting matrix is copied into the upper left quadrant of Table 8. The first diagonal element of Table 8 is 2.00 + 12.274 = 14.274. Note that the elements below the diagonal are not copied in.

12. The diagonal elements in the upper right quadrant of Table 8 are the corresponding elements of row g.

13. Next compute the inverse of the A_1 matrix, postmultiplied by the

diagonal matrix in the upper right quadrant of Table 8. The procedure used is identical with that previously mentioned in connection with computing L_1, and is outlined in ([6], Ch. 21, Sec. 7). Computations for the forward solution are shown in the lower quadrants of Tables 8 and 9. The backward solution is shown in Table 10, which gives the transpose of the desired product matrix in the first six columns.

TABLE 8

Computation of $A_1^{-1}D_{b_2}^{\frac{1}{2}}$, where $A_1 = D_{b_2}^{\frac{1}{2}} RD_{b_2}^{\frac{1}{2}} + \Delta$ Forward Solution

		1A	2A	3A	4A	5A	6A	1B	2B	3B	4B	5B	6B	Check	Σ
	1A	14.274	1.362	1.983	3.639	9.874	7.109	3.653						41.894	41.894
	2A		5.780	.025	3.068	2.426	2.154		2.345					17.160	17.160
	3A			14.137	2.532	1.797	-2.025			3.572				22.021	22.021
	4A				14.446	6.896	5.707				3.545			39.833	39.833
	5A					12.969	8.408					3.543		45.913	45.913
	6A						17.881						3.779	43.013	43.013
		38.241	14.815	18.449	36.288	42.370	39.234	3.653	2.345	3.572	3.545	3.543	3.779		209.834
.0701	1	14.274	1.362	1.983	3.639	9.874	7.109	3.653						41.894	41.894
.1770	2		5.651	-.163	2.722	1.488	1.479	-.347	2.345					13.180	13.175
.0722	3			13.857	2.105	.468	-2.971	-.518	.068	3.572				16.580	16.581
.0841	4				11.886	3.590	3.633	-.686	-1.141	-.543	3.545			20.279	20.284
.2153	5					4.645	2.103	-2.212	-.274	.043	-1.071	3.543		6.768	6.777
.0889	6						11.253	-.627	-.127	.911	-.600	-1.605	3.779	12.969	12.984

TABLE 9

Computation of $A_1^{-1} D_{b_2}^{\frac{1}{2}}$ Continued

	1	2	3	4	5	6	1	2	3	4	5	6	Check	Σ
1	1.000													
2		1.000												
3			1.000											
4				1.000										
5					1.000									
6						1.000								
1	-1.000	-.095	-.139	-.255	-.692	-.498	-.256						-2.935	-2.935
2		-1.000	.029	-.482	-.263	-.262	.061	-.415					-2.331	-2.332
3			-1.000	-.152	-.034	.214	.037	-.005	-.258				-1.197	-1.198
4				-1.000	-.302	-.306	.058	.096	.046	-.298			-1.707	-1.706
5					-1.000	-.453	.476	.059	-.009	.231	-.763		-1.459	-1.459
6						-1.000	.056	.011	-.081	.053	.143	-.336	-1.154	-1.154

TABLE 10

Computation of $A_1^{-1} D_{b_2}^{\frac{1}{2}}$ Backward Solution

Matrix $(A_1^{-1} D_{b_2}^{\frac{1}{2}})$

	1	2	3	4	5	6	1	2	3	4	5	6	Check	
1	.576	.025	-.048	.095	-.451	-.056	-1						0	.000
2	.015	.469	.016	-.076	-.054	-.011		-1					0	-.005
3	-.047	.024	.286	-.062	-.028	.081			-1				0	.004
4	.093	-.115	-.062	.377	-.207	-.053				-1			0	-.006
5	-.438	-.082	-.027	-.206	.828	-.143					-1		0	-.009
6	-.058	-.018	.086	-.057	-.152	.336						-1	0	.001
Σ	.141	.303	.251	.071	-.064	.154								

14. Each column of the matrix obtained in the backward solution now is multiplied by the corresponding element of row g. For the first element in column 1, we have $.576 \times 3.6526 = 2.104$. The resulting matrix, with corresponding off-diagonal elements averaged to make the matrix perfectly symmetrical, is shown in Table 11.

TABLE 11

Computation of $(D_{b_2}^{\frac{1}{2}} A_1^{-1}) D_{b_2}^{\frac{1}{2}}$

	1	2	3	4	5	6	Σ
1	2.104	.057	-.172	.339	-1.599	-.212	.517
2	.057	1.100	.056	-.270	-.192	-.042	.709
3	-.172	.056	1.022	-.220	-.098	.306	.894
4	.339	-.270	-.220	1.337	-.732	-.201	.253
5	-1.599	-.192	-.098	-.732	2.934	-.540	-.227
6	-.212	-.042	.306	-.201	-.540	1.270	.581
Σ	.517	.709	.894	.253	-.227	.581	2.727

TABLE 12

Computation of $L_2' = \left[(D_{b_2}^{\frac{1}{2}} A_1^{-1} D_{b_2}^{\frac{1}{2}})\alpha_c\right]'$

	1	2	3	4	5	6
1	.051	-.053	.073	.051	-.064	.000
2	-.082	.036	.018	.018	.044	.022
3	.003	-.005	-.003	-.009	.024	-.049
4	.131	-.003	-.076	-.062	.054	.052
5	.070	.093	-.108	-.009	-.229	.101
6	-.175	-.043	.100	-.058	.080	-.055
7	.001	-.041	-.008	-.092	.120	-.052
8	-.065	.104	-.085	.086	-.026	.003
9	.076	-.069	.075	.005	.006	.010
10	-.005	-.022	.012	.073	-.014	-.028
Σ	.005	-.003	-.002	.003	-.005	.004
Ck	.005	-.002	-.004	.002	-.004	.003

15. Next compute, by successive columns, the matrix L_2', which is shown in the first ten rows of Table 12. The ith element in the first column of L_2' is the product sum of elements in the first row of Table 11 by the corresponding elements in the ith row of the α_c' matrix in Table 3. For example, the first element in the first column of L_2' is $(2.104)(.023) + (.057)(-.047) + (-.172)(.089) + (.339)(.033) + (-1.599)(-.004) + (-.212)(-.016) = .051$. The second column of L_2' is obtained in the same manner as the first except that the second row of Table 11 is used instead of the first. To obtain the third column of L_2' the third row of Table 11 is used, and so on until the table is completed.

16. Step 8 now is repeated, rows a through f, using the L_2' matrix to obtain a third approximation to the altered test lengths (i.e., a new row f). These

computations are not reproduced here, but the values obtained in row *f* may be seen in the third row of Table 13.

TABLE 13

Differential Prediction: Successive Approximations* to $1'D_b$, for $T_1 = \frac{1}{2}T_0 = 71$

Approx'n		1	2	3	4	5	6	Σ	Value of ∅ for Successive Values of L	
$(0.5)1'D_a$:	1	12.50	4.50	15.00	11.50	7.50	20.00	71.00	L_1	.227
	2	13.34	5.50	12.76	12.57	12.55	14.28	71.00	L_2	.234
	3	13.20	5.31	11.57	12.56	16.06	12.32	71.02	L_3	.236
	4	13.20	5.18	11.04	12.51	17.63	11.44	71.00	L_4	.237
	5	13.26	5.12	10.82	12.56	18.15	11.09	71.00	L_5	.235
	6	13.31	5.12	10.75	12.55	18.37	10.91	71.01	L_6	.238

*The third and subsequent approximations were computed by the procedure described for the general case.

TABLE 14

Differential Prediction: Successive Approximations to $1'D_b$, for $T_1 = \frac{1}{2}T_0 = 71$

From (4, p. 60)

Approx'n		1	2	3	4	5	6	Σ	Value of ∅ for Successive Values of L	
$(0.5)1'D_a$:	1	12.50	4.50	15.00	11.50	7.50	20.00	71.00	L_1	.227
	2	13.34	5.50	12.76	12.57	12.55	14.28	71.00	L_2	.234
	3	13.27	5.32	11.55	12.52	16.05	12.29	71.00	L_3	.235
	4	13.23	5.20	10.98	12.47	17.62	11.51	71.01	L_4	.236
	5	13.31	5.15	10.76	12.46	18.13	11.19	71.00	L_5	.237
	6	13.35	5.12	10.70	12.46	18.37	11.00	71.00	L_6	.236

Computations for the fourth approximation may be summarized as follows: (1) a new row *g* is computed to obtain the square roots of the corresponding values in the new row *f*; (2) steps 9 through 11 are repeated with the new values obtained in row *g* to compute the matrix A_2; (3) steps 12 through 15 are repeated to compute L_3' ; (4) step 16 is repeated to obtain the fourth approximation. Thus, given any approximation, row *g* of step 8 and steps 9 through 16 designate the procedure which may be used with complete generality to compute subsequent approximations to optimal test lengths for differential prediction.

In all, five approximations beyond the first were computed and are summarized in Table 13. Of these, the second was computed by the procedure presented in [4]; approximations three through six were calculated by the procedure described for the general case.

17. Successive indices of differential prediction efficiency, ϕ, are computed as follows:

(a) To obtain ϕ_1 corresponding to the first approximation to the altered test lengths, each element in the L_1' matrix is multiplied by the corresponding element in the α_c' matrix in Table 3, and all products are summed. The resulting value, .227, is found as the first entry in the ϕ column at the extreme right in Table 13.

(b) The value of ϕ_2 is obtained in the same manner except that the L_2' matrix is used.

(c) Subsequent values, ϕ_i , are obtained by using the elements of L_i' and the corresponding elements of α_c' in Table 3.

Table 14 shows the results obtained in [4] by the original procedure, for the same data and with the over-all new testing time also equal to one-half the original time. Comparison of the corresponding values in Tables 13 and 14 indicates results essentially the same for all practical purposes. The largest discrepancy does not exceed one-tenth of a minute, and the increases in ϕ, though small, are comparable within the range of rounding errors. In neither case have computations been carried to the point of complete stabilization of the vector of time allotments. Results, however, appear adequate for practical purposes.

The question may arise as to the stability of the time estimates from sample to sample. The entire problem of significance tests, however, has not yet been touched.

The General Case for Multiple Absolute Prediction

The computational procedure presented in [5] for obtaining optimal test length for multiple absolute prediction consists of the same sequence of operations as that given in [4], the difference being that in [5] the matrix of validity coefficients is used, whereas in [4] these coefficients in deviation form for each test were required.

Similarly, the sequence of operations for the general case for multiple absolute prediction is the same as that presented above, the difference being that the matrix r_c is used, whereas α_c was required above. Instead of presenting a numerical example in detail for the general case for multiple absolute prediction, here only the procedural steps which differ from those described above will be indicated. Namely:

Step 1 is omitted.

In step designated 6–7, the r_c matrix is used instead of matrix α_c .

In steps 15 and 17, the r_c' matrix is used instead of matrix α_c' . The above distinctions assume, as was assumed for the general case for differential prediction, that the second approximation to altered time allotments was computed by the original method.

The series of approximations to optimal test lengths for maximum absolute prediction, shown in Table 15, was obtained with the same original data as the series in the previous example, but with the over-all testing time taken as unchanged. Hence the original test lengths were taken as the first approximation.

To demonstrate that the procedure developed for the general case yields the same results as the procedure presented in [5], the procedure described for for general case was employed immediately. The square roots of the original test lengths were found at once, as designated by step 8, row *g* of the preceding example. Further procedural steps followed the directions given in the steps subsequent to step 8, row *g*, with the exception, of course, that in steps 15 and 17, the r_c' matrix was used instead of matrix α_c' .

TABLE 15·

Absolute Prediction: Successive Approximations* to $1'D_b$, for $T_1 = T_0 = 142$

Approx'n		1	2	3	4	5	6	Σ	Value of λ for Successive Values of L	
$(1.0)1'D_a$:	1	25.00	9.00	30.00	23.00	15.00	40.00	142.00	L_1	2.203
	2	32.53	10.02	10.50	21.40	18.65	48.89	141.99	L_2	2.229
	3	32.82	9.67	8.30	20.31	21.60	49.30	142.00	L_3	2.230
	4	32.79	9.53	7.64	19.96	23.03	49.05	142.00	L_4	2.230

*The second, third and fourth approximations were computed by the procedure described for the general case.

TABLE 16

Absolute Prediction: Successive Approximations to $1'D_b$, for $T_1 = T_0 = 142$

From (5, p. 120)

Approx'n		1	2	3	4	5	6	Σ	Value of λ for Successive Values of L	
$(1.0)1'D_a$:	1	25.00	9.00	30.00	23.00	15.00	40.00	142.00	L_1	2.203
	2	32.54	10.00	10.45	21.42	18.66	48.93	142.00	L_2	2.230
	3	32.87	9.70	8.21	20.21	21.61	49.40	142.00	L_3	2.234
	4	32.76	9.52	7.57	19.99	23.08	49.08	142.00	L_4	2.232

Three approximations beyond the original values were computed. These, with the corresponding values of the index of multiple absolute prediction efficiency, λ, are shown in Table 15. The corresponding results obtained by the original method are found in Table 16. A comparison of the two tables shows no discrepancy in the entire series greater than one-tenth of a minute, and no difference between the corresponding values of λ beyond those of rounding errors.

Mathematical Derivation

The General Case for Differential Prediction

The mathematical rationale presented in [4] provides a solution for obtaining optimal test lengths by means of a series of approximations. The formulas derived are not applicable, however, in the event that the altered time allotment for any test approaches zero. The derivation which follows consists in developing, from the computational equations presented in [4], formulas which do not involve reciprocals of the altered time allotments, and which, consequently, provide a solution such that its applicability is perfectly general. Using the notation of [4], let

n $\equiv$ the number of predictors,
N $\equiv$ the number of criteria,
r $\equiv$ the $(n \times n)$ matrix of intercorrelations of tests of original lengths,
r_c $\equiv$ the $(n \times N)$ matrix of validity coefficients for the tests of original lengths,
D_a $\equiv$ the $(n \times n)$ diagonal matrix of original test lengths,
D_b $\equiv$ the $(n \times n)$ diagonal matrix of altered test lengths,
$D_{r_{ii}}$ $\equiv$ the $(n \times n)$ diagonal matrix of reliability coefficients for the tests of original lengths.

As in [4], define

$$R \equiv r - (I - D_{r_{ii}}),$$

$$\alpha_c \equiv r_c\left(I - \frac{11'}{N}\right),$$

$$\Delta \equiv D_a(I - D_{r_{ii}}),$$

and again state the constraining condition, $T = 1'D_b1$, where 1 is a vector of unities.

Start with equations (43) and (44) of [4], namely, the equations from which the formulas for the iterative solution for D_b were derived. These are, respectively,

$$D_b = \frac{(D_{LL'} \cdot \Delta)^{1/2} T}{1'(D_{LL'} \cdot \Delta)^{1/2} 1}, \tag{1}$$

and

$$L = (R + \Delta D_b^{-1})^{-1}\alpha_c \; , \tag{2}$$

where $D_{LL'}$ is a diagonal matrix whose non-zero elements are the diagonal elements of LL'. Equation (2) may also be expressed as

$$L = (D_b^{-1/2}D_b^{1/2}RD_b^{1/2}D_b^{-1/2} + D_b^{-1/2}\Delta D_b^{-1/2})^{-1}\alpha_c \; , \tag{3}$$

or equivalently, as

$$L = [D_b^{-1/2}(D_b^{1/2}RD_b^{1/2} + \Delta)D_b^{-1/2}]^{-1}\alpha_c \; , \tag{4}$$

or finally, as

$$L = D_b^{1/2}(D_b^{1/2}RD_b^{1/2} + \Delta)^{-1}D_b^{1/2}\alpha_c \; , \tag{5}$$

an equation which involves no negative powers of D_b .

Let

$$L_i = D_{b_i}^{1/2}(D_{b_i}^{1/2}RD_{b_i}^{1/2} + \Delta)^{-1}D_{b_i}^{1/2}\alpha_c \; , \tag{6}$$

where

$$D_{b_1} = \frac{T}{1'D_a1} D_a \; , \tag{7}$$

and

$$D_{b_{i+1}} = \frac{(D_{L_iL'_i}\Delta)^{1/2}T}{1'(D_{L_iL'_i}\Delta)^{1/2}1} . \tag{8}$$

The first approximation to D_b is indicated by (7). The second and all subsequent approximations to D_b may be obtained by an iterative procedure based on (6) and (8). In this manner, successive approximations to L_i and $D_{b_{i+1}}$ may be computed until D_b stabilizes satisfactorily.

The General Case for Multiple Absolute Prediction

By an analogous development, it can be shown that for the general case for multiple absolute prediction the formula for successive approximations to L is

$$L_i = D_{b_i}^{1/2}(D_{b_i}^{1/2}RD_{b_i}^{1/2} + \Delta)^{-1}D_{b_i}^{1/2}r_c \; , \tag{9}$$

and that the formulas for obtaining the first and subsequent approximations to D_b are identical with (7) and (8) above.

REFERENCES

[1] Horst, P. Determination of optimal test length to maximize the multiple correlation. *Psychometrika*, 1949, **14**, 79-88.

[2] Horst, P. A technique for the development of a differential prediction battery. *Psychol. Monogr.* 1954, **68**, No. 9 (Whole No. 380).

[3] Horst, P. A technique for the development of a multiple absolute prediction battery. *Psychol. Monogr.* 1955, **69**, No. 5 (Whole No. 390).

[4] Horst, P. Optimal test length for maximum differential prediction. *Psychometrika*, 1956, 21, 51-66.

[5] Horst, P. and MacEwan, Charlotte. Optimal test length for maximum absolute prediction. *Psychometrika*, 1956, **21**, 111-124.

[6] Horst, P. Servant of the human sciences. Unpublished manuscript. Division of Counseling and Testing Services, Univ. of Washington, May 1953.

Manuscript received 3/18/57

PSYCHOMETRIKA—VOL. 22, NO. 4
DECEMBER, 1957

STIMULUS AND RESPONSE GENERALIZATION: A STOCHASTIC MODEL RELATING GENERALIZATION TO DISTANCE IN PSYCHOLOGICAL SPACE*

ROGER N. SHEPARD

NAVAL RESEARCH LABORATORY†

A mathematical model is developed in an attempt to relate errors in multiple stimulus-response situations to psychological inter-stimulus and inter response distances. The fundamental assumptions are (*a*) that the stimulus and response confusions go on independently of each other, (*b*) that the probability of a stimulus confusion is an exponential decay function of the psychological distance between the stimuli, and (*c*) that the probability of a response confusion is an exponential decay function of the psychological distance between the responses. The problem of the operational definition of psychological distance is considered in some detail.

Stochastic models for learning have been developed by Estes [8], by Bush and Mosteller [6], and others. With the exception of a few investigations confined to the stimulus side of the learning process, such as that by Bush and Mosteller [5], however, these models have not been extensively applied to generalization phenomena. This paper, in using the notion of *psychological distance,* approaches the generalization problem from a somewhat different direction.

Consideration will be restricted to situations in which a number of responses are discriminatively attached to a number of stimuli by consistent application of differential reinforcement. More precisely, the learning process will be supposed to conform to the following rules: (*a*) On any given trial a single stimulus is presented at random from a set of N stimuli. (*b*) On each trial the subject is constrained to a fixed set of N responses. (*c*) For any given subject, there is a prevailing one-to-one assignment of the N responses to the N stimuli arbitrarily determined in advance such that a certain reinforcing operation (e.g., the word "correct") is applied if and only if the presentation of a stimulus is followed by the occurrence of its assigned response. The present model, however, is concerned not with the learning process per se but with the pattern of generalizations exhibited at any one given

*This paper is based upon the theoretical sections of a Ph.D. dissertation submitted to the Graduate School of Yale University and upon subsequent modification carried out on a National Academy of Sciences-National Research Council Postdoctoral Associateship at the Naval Research Laboratory. The author is particularly indebted to Drs. C. I. Hovland, R. P. Abelson, and B. S. Rosner for their generous advice and support. Helpful criticisms have also been contributed by Drs. G. A. Miller, F. A. Logan, W. D. Garvey, J. G. Holland, and H. Glaser.

†Now at Psychological Laboratories, Harvard University.

stage of learning. Furthermore, as an approximation for any given set of stimuli or responses, all subjects are assumed to generalize according to the same pattern.

Ordinarily there is no necessary or natural correspondence between the stimuli and responses and, indeed, different assignments may be set up for different subjects. It is convenient, therefore, to have a way of referring to the response which has been assigned, for a given subject m, to a given stimulus, S_i , without having to ask just which one of the N responses that may be. Accordingly, the following definitions are introduced:

$S_i \equiv$ the ith of the N stimuli $S_1 , S_2 , \cdots , S_N$;
$R_i \equiv$ the ith of the N responses $R_1 , R_2 , \cdots , R_N$;
$R_{(i),m} \equiv$ the response assigned to S_i for subject m;
$S_{(i),m} \equiv$ the stimulus to which R_i is assigned for subject m.

Thus the set of all stimulus-response sequences, for any subject m, divides into (a) the subset of reinforced sequences of the form $S_i \rightarrow R_{(i),m}$ and (b) the subset of nonreinforced sequences of the form $S_i \rightarrow R_{(k),m}$ with $i \neq k$.

At any given stage of learning there will be, for every stimulus and every response, a probability that the one will be followed by the other. These probabilities are designated as follows:

$P_{ik,m} = P[R_k \mid S_i]_m \equiv$ the conditional probability of R_k , given S_i , for subject m.

$P_{i(k),m}$, $P_{(i)k,m}$, and $P_{(i)(k),m}$ are defined in an analogous manner. Thus

$P_{i(k),m} = P[R_{(k),m} \mid S_i]_m \equiv$ the conditional probability of $R_{(k),m}$, given S_i , for subject m.

The responses are partitioned so as to be mutually exclusive and exhaustive. The conditional probabilities, therefore, satisfy the requirements

$$\sum_k P_{i(k),m} = 1, \qquad P_{i(k),m} \geq 0. \tag{1}$$

If the probabilities $P_{i(i),m}$ increase with continued application of the reinforcing operation, there must result a decrease in some $P_{i(k),m}$ with $i \neq k$. Although it is known that the probabilities of the various incorrect responses, called generalization errors, do not in general decay at the same rate, little advance has been made towards the quantitative understanding of this aspect of the learning process. About all one has to go on is the qualitative observation that, at a given stage of learning, the probability, $P_{i(k),m}$, decreases both with the dissimilarity of S_i and S_k , and with the dissimilarity of $R_{(i),m}$ and $R_{(k),m}$.

The Reduction of the S-R Process to an S-S and R-R Process

An error in which the response assigned to a stimulus, S_j , follows the presentation of another, S_i , (as in B of Fig. 1) may be viewed as comprising

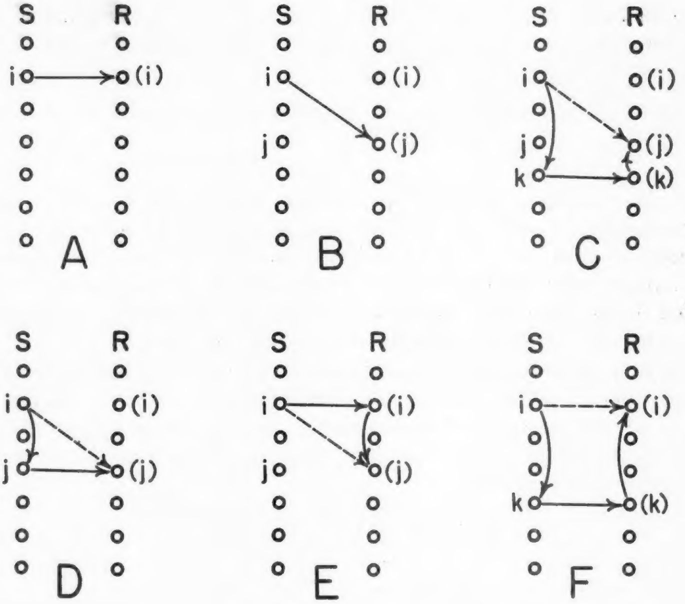

FIGURE 1

Different ways of conceptualizing an *S-R* sequence as generated by an *S-S* and an *R-R* sequence. The circles on the left stand for the different stimuli and the circles on the right for their assigned responses. In A, for example, S_i was presented and followed by its assigned response, $R_{(i),m}$. In B, however, S_i was followed by an incorrect response, $R_{(j),m}$.

two events as illustrated in C. It may be that the subject confused two stimuli: when S_i was presented, it was taken to be S_k . Now, if S_k is taken to be the stimulus, the response which should ensue is $R_{(k),m}$. Suppose, however, that the subject also confused two responses; whereas the tendency was to make $R_{(k),m}$, the response actually made (according to the external criteria) was $R_{(j),m}$. In this way an *S-R* transition may be analyzed into an *S-S* transition and an *R-R* transition. Alternatively, there may be a stimulus confusion without any response confusion (D), a response confusion without any stimulus confusion (E), or both stimulus and response confusions which so counteract each other that the correct response is made (F).

This analysis suggests the following additional definitions:

$P^S_{ik} = P[S'_k \mid S_i] \equiv$ the conditional probability that S_i , when presented, will be taken to be S_k .

$P^R_{ik} = P[R_k \mid R'_i] \equiv$ the conditional probability that R_k will be made in place of R_i .

The term introduced in the second definition is also the probability that, when the stimulus is taken to be S_i, R_k will follow, since the model is set up so that

$$P[R_k' \mid S_i'] = \begin{cases} 1, & \text{if} \quad i = k, \\ 0, & \quad \text{otherwise.} \end{cases}$$

Thus, the conditional probabilities are treated as if the subject (*a*) knows the connection between any stimulus and its assigned response but (*b*) still makes errors owing to a certain inability to identify the stimulus or reproduce the response with sufficient accuracy (see Fig. 1). The analysis, then, is applied to the confusions among the stimuli and among the responses but not to the associations between the stimuli and the responses.

Strictly speaking, the *S-S* and *R-R* transition probabilities pertain to a short interval of time during the learning process so that a stable state may be assumed to exist. Over extended periods the probabilities will change owing to the effects of reinforcement. At any given time, however, they must satisfy the conditions

$$\sum_k P_{ik}^S = 1, \qquad P_{ik}^S \geq 0, \tag{2}$$

$$\sum_k P_{ik}^R = 1, \qquad P_{ik}^R \geq 0. \tag{3}$$

The fundamental assumption which will be made here is that, at each stage of learning, the response confusions are independent of the stimulus confusions, or, more explicitly, that $P_{(i)(k),m}^R$ does not depend upon which stimulus was presented (and taken to be S_i) on the trial considered. This assumption implies that

$$P_{i(j),m} = \sum_k P_{ik}^S \cdot P_{(k)(j),m}^R \,. \tag{4}$$

Since the indices i, j, and k are allowed to range over the values $1, 2, \cdots, N$, (4) corresponds to the defining equation for matrix multiplication, so that

$$\mathbf{P}_{S(R),m} = \mathbf{P}_{SS} \cdot \mathbf{P}_{(R)(R),m} \tag{5}$$

where, for example,

$$\mathbf{P}_{S(R),m} = \begin{bmatrix} P_{1(1),m} & P_{1(2),m} & \cdots & P_{1(N),m} \\ P_{2(1),m} & P_{2(2),m} & \cdots & P_{2(N),m} \\ \vdots & \vdots & & \vdots \\ P_{N(1),m} & P_{N(2),m} & \cdots & P_{N(N),m} \end{bmatrix}. \tag{6}$$

(Dr. Burton S. Rosner has independently proposed essentially the same treatment for the *S-R* process.)

As the form of the definitions suggests, there is an *S-R* symmetry in the notation such that the content of (4) can be set down in an alternative form giving the probability, when $S_{(i)}$ is presented to subject m, that R_j will ensue.

$$P_{(i)j,m} = \sum_k P^S_{(i)(k),m} \cdot P^R_{kj} \, . \tag{7}$$

The matrix representation becomes

$$\mathbf{P}_{(S)R,m} = \mathbf{P}_{(S)(S),m} \cdot \mathbf{P}_{RR} \, . \tag{8}$$

Equivalent equations (5) and (8) can be reduced to a single form without parentheses around the subscripts. In order to do this, it is convenient to introduce permutation matrices, $\mathbf{J}_m$, with elements $J_{ik,m}$, such that

$$J_{ik,m} = \begin{cases} 1, & \text{if } R_k \text{ is assigned to } S_i \text{ for subject } m, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

By carrying out the indicated multiplications, the following identities may be verified.

$$\begin{aligned} \mathbf{P}_{xS,m} &= \mathbf{P}_{x(S),m} \cdot \mathbf{J}'_m \, , \\ \mathbf{P}_{xR,m} &= \mathbf{P}_{x(R),m} \cdot \mathbf{J}_m \, , \\ \mathbf{P}_{Sx,m} &= \mathbf{J}_m \cdot \mathbf{P}_{(S)x,m} \, , \\ \mathbf{P}_{Rx,m} &= \mathbf{J}'_m \cdot \mathbf{P}_{(R)x,m} \, , \\ \mathbf{J}_m \cdot \mathbf{J}'_m &= \mathbf{J}'_m \cdot \mathbf{J}_m = \mathbf{I}. \end{aligned} \tag{10}$$

Here, the subscript x stands for any of the symbols S, (S), R, or (R); $\mathbf{J}'_m$ is the transpose of $\mathbf{J}_m$; and $\mathbf{I}$ is the identity matrix.

Using these relations, (5) and (8) can be brought into the form

$$\mathbf{P}_{SR,m} = \mathbf{P}_{SS} \cdot \mathbf{J}_m \cdot \mathbf{P}_{RR} \, , \tag{11}$$

where $\mathbf{P}_{SR,m}$ is the matrix of *S-R* transition probabilities $P_{ij,m}$.

Now, although $\mathbf{J}_m$ is known and $\mathbf{P}_{SR,m}$ can be estimated from the experimental data, neither $\mathbf{P}_{SS}$ nor $\mathbf{P}_{RR}$ can be directly determined. It might be thought, however, that by assigning the responses to the stimuli in a different way for each subject, the influence of the response confusions could be counterbalanced out of (11) so that $\mathbf{P}_{SS}$ could be solved for in terms of $\mathbf{P}_{SR}$. Suppose, then, that there are M subjects, each with a different assignment, so that, over the set of all M assignments, every pair of responses is assigned to each pair of stimuli the same number of times in both of the two possible orders.

Using the simplified notations

$$\mathbf{P}_{S(R)} = \frac{1}{M} \sum_m \mathbf{P}_{S(R),m} , \tag{12}$$

$$\mathbf{P}_{(R)(R)} = \frac{1}{M} \sum_m \mathbf{P}_{(R)(R),m} , \tag{13}$$

and assuming that all the subjects are essentially alike so that $\mathbf{P}_{SS}$ is independent of m, equation (5) may be averaged over all M subjects to yield

$$\mathbf{P}_{S(R)} = \mathbf{P}_{SS} \cdot \mathbf{P}_{(R)(R)} .$$

Postmultiplying through by $\mathbf{P}^{-1}_{(R)(R)}$,

$$\mathbf{P}_{SS} = \mathbf{P}_{S(R)} \cdot \mathbf{P}^{-1}_{(R)(R)} . \tag{14}$$

The assumption that all subjects tend to confuse stimuli in accordance with the same pattern is analogous to the similar assumption made in order to pool data from different subjects in psychological scaling procedures. This assumption is probably correct only as a first approximation since the tendency to confuse any particular pair of stimuli probably depends to some degree upon the history of discrimination learning associated with that pair.

In order to evaluate the inverse R-R matrix, it may be noted from (10) that

$$\mathbf{P}_{(R)(R),m} = \mathbf{J}_m \cdot \mathbf{P}_{RR} \cdot \mathbf{J}'_m , \tag{15}$$

where the assignments are so chosen that the matrices $\mathbf{J}_m$ and $\mathbf{J}'_m$ select, for each nondiagonal cell of $\mathbf{P}_{(R)(R)}$, elements P^R_{ik} $(i \neq k)$ from every nondiagonal cell of $\mathbf{P}_{RR}$ an equal number of times, as m ranges from 1 to M. Likewise, for each diagonal cell of $\mathbf{P}_{(R)(R)}$, elements P^R_{ii} will be equally selected from each diagonal cell of $\mathbf{P}_{RR}$. Thus, by definition of the matrix elements $J_{ik,m}$ and $J'_{ik,m}$, averaging over all assignments insures that

$$\begin{aligned} \frac{1}{M} \sum_m P^R_{\substack{(i)(k),m \\ (i \neq k)}} &= \frac{1}{M} \sum_m \sum_{\substack{g \\ (g \neq h)}} \sum_h J_{ig,m} \cdot P^R_{gh} \cdot J'_{hk,m} \\ &= \frac{1}{N(N-1)} \sum_{g} \sum_{\substack{h \\ (g \neq h)}} P^R_{gh} = P^R, \\ \frac{1}{M} \sum_m P^R_{(i)(i),m} &= \frac{1}{M} \sum_m \sum_g J_{ig,m} \cdot P^R_{gg} \cdot J'_{gi,m} \\ &= \frac{1}{N} \sum_g P^R_{gg} = Q^R, \end{aligned} \tag{16}$$

where P^R is the mean probability that two different responses are confused and Q^R is the mean probability that a response is not confused with any other. Since the total probability must be conserved,

$$Q^R + (N - 1)\cdot P^R = 1. \tag{17}$$

One way of understanding the operations represented in (16) is to multiply, e.g., an arbitrary 3×3 matrix (for $\mathbf{P}_{RR}$) by all six possible 3×3 permutation matrices and their transposes as indicated in (15). The sum of these products will contain equal nondiagonal elements, P^R, and equal diagonal elements, Q^R, as required. A minimum set of permutation matrices having the necessary properties for $N = 9$ is given in the appendix.

Equation (14) may thus be written in the form

$$\mathbf{P}_{SS} = \mathbf{P}_{S(R)}\cdot\begin{bmatrix} Q^R & P^R & \cdots & P^R \\ P^R & Q^R & \cdots & P^R \\ \vdots & \vdots & & \vdots \\ P^R & P^R & \cdots & Q^R \end{bmatrix}^{-1}. \tag{18}$$

But the inverse matrix has a simple representation such that

$$\mathbf{P}_{SS} = \mathbf{P}_{S(R)}\cdot\left[\frac{1}{Q^R - P^R}\right]\cdot\begin{bmatrix} (1 - P^R) & -P^R & \cdots & -P^R \\ -P^R & (1 - P^R) & \cdots & -P^R \\ \vdots & \vdots & & \vdots \\ -P^R & -P^R & \cdots & (1 - P^R) \end{bmatrix}, \tag{19}$$

as may be verified by using (17) and showing that the product of the original matrix (with elements P^R and Q^R) and its inverse representation in (19) yields the identity matrix.

Expanding with respect to the general term in (19), the probability that S_i will be taken to be S_k is

$$P^S_{ik} = \frac{1 - P^R}{Q^R - P^R} P_{i(k)} - \frac{P^R}{Q^R - P^R} \sum_{\substack{j \\ (j \neq k)}} P_{i(j)} .$$

Using (1) and (17), this may be reduced to

$$P^S_{ik} = \frac{P_{i(k)} - P^R}{1 - NP^R}. \tag{20}$$

Thus, although the response confusions are not entirely eliminated by employing different *S-R* assignments, they are consolidated in the single parameter P^R. In principle this parameter could be empirically estimated in special cases (e.g., with unidimensional stimuli) by extrapolating a fitted *S-R* transition probability function (for pairs of stimuli) in the direction of increasing stimulus dissimilarity. For, by (20),

$$\text{as}\quad P^S_{ik} \rightarrow 0, \qquad P_{i(k)} \rightarrow P^R.$$

In practice, if the responses are highly distinctive so that P^R is close to zero, the probabilities $P_{i(k)}$ can be taken as estimates of the probabilities P^S_{ik} with less *R-R* probability contamination than would be possible without the counterbalancing technique. The reason for this will appear in the discussion of estimation procedures.

If the individual *R-R* transition probabilities are desired, (8) may be averaged over all M assignments. Employing arguments analogous to those developed before, the probability of an *R-R* transition is found to be

$$P^R_{ik} = \frac{P_{(i)k} - P^S}{1 - NP^S}. \tag{21}$$

This is the inverse of (20) in that P^S denotes the mean transition probability taken over all pairs of stimuli.

The utility of reducing the *S-R* transitions to *S-S* and *R-R* transitions can now be ascribed to the consequent increase in predictive power of the model. If the *S-S* probability matrices have been determined (in H experiments) for each of H sets of N stimuli and if the *R-R* matrices have been determined (in H further experiments) for each of H sets of N responses, the total number of experiments for which the *S-R* probability matrices can be predicted is $N!\cdot H^2$. For, returning to (11), there are $N!$ distinct matrices which are substitutable for $\mathbf{J}_m$ (each corresponding to a different *S-R* assignment with one set of stimuli and one set of responses) and there are H^2 distinct pairs consisting of one set of stimuli and one set of responses. The ratio of the number of experiments for which prediction can be made to the number already carried out is $N!\cdot H/2$. In contrast to this, if the *S-R* probabilities are regarded as irreducible, predictions could be made only to replications of experiments already carried out, and the ratio just considered could never exceed unity.

The Characterization of the S-S and R-R Processes in Terms of Psychological Inter stimulus and Inter response Distances

In the preceding section the *S-R* process was reduced to *S-S* and *R-R* processes which, in turn, were characterized by matrices of *S-S* and *R-R* transition probabilities. The purpose of the present section will be to reduce these matrices, each with $N(N - 1)$ independent probabilities, to sets of fewer than $N(N - 1)$ quantities. Such a reduction is suggested by the possibility that some simple relation exists between P^S_{ik} and P^S_{ki} of the *S-S* matrix, and between P^R_{ik} and P^R_{ki} of the *R-R* matrix. This, in turn, appears plausible if the probability of confusing two stimuli (or responses) is some function of the dissimilarity between them so that, say, P^S_{ik} and P^S_{ki} will increase or decrease together as the dissimilarity between S_i and S_k is made, respectively, smaller or larger.

Instead of formally introducing the notion of dissimilarity, it is preferable to define the concept of distance, which has the advantage of a rigorous mathematical interpretation. Explicitly, a set of distances, D_{ik}, defined for all pairs of elements, S_i and S_k, is any collection of numbers satisfying, for every S_i, S_j, and S_k, the following requirements called metric axioms ([2], pp. 5–16, [16], pp. 118–119):

$$D_{ik} = 0, \quad \text{if} \quad i = k, \tag{22}$$

$$D_{ik} = D_{ki}, \tag{23}$$

$$D_{ik} + D_{kj} \geq D_{ij}. \tag{24}$$

When speaking of the distance between S_i and S_k, the symbol D^S_{ik} will be used. Similarly the distance between R_i and R_k will be distinguished by the symbol D^R_{ik}. Any set of elements for which a distance function satisfying the metric axioms has been defined is called a metric space. The space may be called a physical or a psychological space depending upon whether the distances are determined from physical or psychological data. (An example of a physical space is the set of sinusoidal tones with

$$D_{ik} = [(f_i - f_k)^2 + (a_i - a_k)^2]^{1/2},$$

where f_i is the frequency and a_i the amplitude of tone S_i. That this definition satisfies axioms (22) and (23) is immediately clear. The satisfaction of (24) follows from the inequality of Schwarz. For a review of some psychological measures which could presumably be used to construct a psychological space for this same set of tones, see Messick [17].)

It will be assumed that there exists some function, f, such that P^S_{ik} is proportional to $f(D^S_{ik})$. The factor of proportionality must depend upon i for, although the average distance of S_i from the other stimuli in the learning situation may be large or small, the probability of transition from S_i to S_k summed over all k must be conserved by equation (2). Thus the relation may be set down in the preliminary form

$$P^S_{ik} = d_i \cdot f(D^S_{ik}), \tag{25}$$

where d_i is a constant associated with S_i, and where D^S_{ik} satisfies the metric axioms. Summing over all k,

$$\sum_k P^S_{ik} = 1 = d_i \cdot \sum_k f(D^S_{ik}).$$

Solving for d_i, it is immediately found that

$$P^S_{ik} = \frac{f(D^S_{ik})}{\sum\limits_h f(D^S_{ih})}. \tag{26}$$

At this point some decision must be reached concerning the nature of the function f. This, of course, is one way of formulating the problem traditionally investigated in studies of stimulus generalization. The independent variable in such studies is some measure of stimulus dissimilarity, and the dependent variable is some measure of stimulus confusability (like the probability that the response, reinforced to one stimulus, will occur to the other).

Now, although these studies lend support to the conjecture that f is a continuous monotonically decreasing function, attempts to specify it with greater precision have not led to any consistent picture ([14], pp. 616–617, [26], pp. 577–579). This may be a consequence, at least in part, of the variety of independent measures employed. The most frequent measures of dissimilarity which have been used are distance on a physical scale and number of just noticeable differences, JNDs, separating two stimuli. However, there are theoretical objections to either of these measures.

That psychological distance or confusion probability is not an invariant function of physical distance is now well known. Some investigators, though, have supposed that the summation of JNDs provides the kind of measure required ([15], pp. 183–225). Unfortunately, in order to sum JNDs between two stimuli, this summation must be carried out along some path between these stimuli. But the resulting sum will be invariant and, therefore, possess fundamental significance only if this path is a least path, that is, yields a shortest distance (in psychological space) between the two stimuli. One cannot presume, in arbitrarily holding certain physical parameters constant (as is ordinarily done in the summation of JNDs), that the summation is constrained thereby to a shortest path (or geodesic) in psychological space, even though it is, of course, confined to a shortest path (or straight line) in physical space. Indeed, given any particular summation, there is no way of ascertaining whether it was or was not carried out over a least path.

These considerations lead one to look for some way of estimating the psychological distance between two stimuli without depending either upon physical scales or upon any arbitrary path of integration. One possibility, suggested by Gulliksen and Wolfle [11], is to use a judgmental procedure in which subjects directly estimate the similarity of stimuli. Indeed, Plotkin [22] and, later, Attneave [1] in their studies of stimulus generalization in paired-associates learning have used techniques of this kind. Although they were able to demonstrate a positive correlation between confusion frequency and judged similarity of stimuli, the exact form of the relation between these variables was not pursued.

More recently, a number of multidimensional scaling methods have been developed which make possible the determination of a set of interstimulus distances solely on the basis of similarity judgments [17]. Thus one might now extend the kind of approach proposed by Gulliksen and Wolfle to the quantitative study of generalization in paired-associates learning. However, beyond the fact that these judgmental methods are limited in application to mature human subjects, there appears to be no readily available means for interpretation of the dependent variables of these scaling procedures within the framework of existing behavior theory. Furthermore these methods have not, as yet, been extended to the response domain.

To maintain the integrity of the present approach, it seems desirable to avoid the use of methods which essentially fall outside the scope of the behavior model to be constructed. Instead of starting with an arbitrary measure of psychological distance in order to discover the relation between this measure and consequent stimulus confusion probabilities (the traditional approach to the generalization problem), the present strategy will be to begin with the confusion probabilities themselves and then, proceeding in the reverse direction, to discover a function, f, which will transform these probabilities into measures satisfying the metric axioms.

Actually, there are many functions having the necessary properties. This is because the so-called triangle axiom given in inequality (24) is not particularly stringent. However, that requirement can be usefully strengthened by making the reasonable assumption that physical space can be mapped into psychological space by a transformation which is not only continuous but also has continuous first partial derivatives. Such a transformation carries any straight line in physical space into a smooth (differentiable) curve in psychological space.

The importance of this assumption derives from the fact that a segment of a differentiable curve approximates more and more closely a segment of a straight line as the two segments are made shorter and shorter. In the limit, for three stimuli, S_i , S_j , and S_k , such that S_k is between S_i and S_j on a single physical dimension, axiom (24) should go over into

$$D_{ik}^{S} + D_{kj}^{S} = D_{ij}^{S} , \tag{27}$$

provided that S_i and S_j are sufficiently close together in physical space.

It is possible to demonstrate (by introducing further assumptions) that an exponential decay form can be deduced for the function f so that (27) will be satisfied for any three properly chosen stimuli. In order to maintain the continuity of the present argument, however, it will be taken as primitive that f is an exponential decay function. The justification for this choice will have to depend upon the empirical results which follow from its use. It might be noted, however, that, apparently largely on the basis of Hovland's results [13], Hull postulated an exponential decay function ([15], pp. 183–225). Other generalization studies have also obtained data roughly consistent with this assumption [3, 9, 10, 12, 19, 23]. In any case, the exponential function is perhaps the simplest function with the desirable behavior that, as its argument ranges from zero over all positive bounds, it subsides asymptotically from a finite value towards zero.

Substituting an exponential decay function for f, (26) becomes

$$P_{ik}^{S} = \frac{\exp(-D_{ik}^{S})}{\sum_{h} \exp(-D_{ih}^{S})}. \tag{28}$$

Since psychological distance is symmetric in accordance with (23) and since the psychological distance between any stimulus and itself must be zero by (22), the entire matrix of transition probabilities may now be reconstructed on the basis of just $N(N - 1)/2$ distances. Thus only half of the degrees of freedom in the probability matrix may really be free in the theoretical sense.

The analogue of (28) is, for the response process,

$$P_{ik}^{R} = \frac{\exp(-D_{ik}^{R})}{\sum_{h} \exp(-D_{ih}^{R})}. \tag{29}$$

The choice of the exponential function in (29) has little precedence since the response generalization function has been investigated in only a limited number of studies [7, 20]. The particular choice is made on the basis of the same arguments assembled in support of the selection of that function in relation to the stimulus process. In addition, the symmetry of the model suggests that the same function may apply in both cases.

The Introduction of the Stimulus and Response Weights

After affecting the reductions of the last section, it must be acknowledged that they were based on the questionable assumption that the psychological distance from S_i to S_k is always identical to the psychological distance from S_k to S_i. It has, for instance, long been known that unfamiliar stimuli tend to be mistaken for familiar stimuli. Thus, under brief exposure, *downwark* may be seen as *downright*, whereas the reverse seldom occurs ([21], p. 360). One might therefore suppose that the distance from *downwark* to *downright* is considerably less than the distance in the reverse direction.

Now, although it is possible to construct a consistent distance geometry even if the symmetry requirement is dropped from the metric axioms ([4], pp. 3–4), to do so here would be to relinquish that possibility which provided the impetus for introducing the distance notion in the first place, that is, the possibility of completely characterizing the S-S probability matrix in terms of a substantially reduced number of quantities.

However, it may be that apparent violations of distance symmetry can always be traced to some factor, like familiarity, which pertains to individual (rather than to pairs of) stimuli and which has the consequence that the presentation of one stimulus leads to the perception of another more frequently than would be expected knowing the probability of perceptual distortion in the opposite direction. With each S_k, then, there may be associated a weight, W_k^S, such that, if S_i is presented, the probability of perceiving S_k is proportional to W_k^S. Equation (28) will then assume the modified form

$$P_{ik}^S = \frac{W_k^S \exp\,(-\,D_{ik}^S)}{\sum_h W_h^S \exp\,(-\,D_{ih}^S)}. \tag{30}$$

The introduction of the term W_k^S provides for both the redundancy in the probability matrix and, presumably, any asymmetry among the stimulus similarities. In particular it is assumed that the entire S-S transition probability matrix can now be reconstructed on the basis of $N(N - 1)/2$ independent distances together with N weights. The ratio of the number of independent quantities reconstructed to the number used in the reconstruction, therefore, can be shown to be $2(N - 1)/(N + 1)$.

Postulating that the response process also involves, for every response R_k, a weight or response preference W_k^R, (29) becomes

$$P_{ik}^R = \frac{W_k^R \exp\,(-\,D_{ik}^R)}{\sum_h W_h^R \exp\,(-\,D_{ih}^R)}. \tag{31}$$

In order to estimate the stimulus weights from the S-S probabilities, (30) may be taken as a starting point. The probability that S_i will be correctly perceived is

$$P_{ii}^S = \frac{W_i^S \exp\,(-\,D_{ii}^S)}{\sum_h W_h^S \exp\,(-\,D_{ih}^S)}. \tag{32}$$

Since the weights occur in both the numerator and denominator, they can be viewed as containing some arbitrary multiplicative factor which may be chosen, for convenience, so that

$$\frac{1}{N} \sum_k W_k^S = 1. \tag{33}$$

Now from (22) it follows that $\exp(-D_{ii}^S) = 1$. Whence

$$\frac{P_{ik}^S}{P_{ii}^S} = \frac{W_k^S}{W_i^S} \cdot \exp\,(-\,D_{ik}^S). \tag{34}$$

To eliminate the distance term one has simply to form the product

$$\left(\frac{P_{ik}^S}{P_{ii}^S}\right)^{1/2} \cdot \left(\frac{P_{ki}^S}{P_{kk}^S}\right)^{-1/2} = \frac{W_k^S}{W_i^S}. \tag{35}$$

If this equation is summed over all i, the weights may be obtained except for a factor, $\sum_i (1/W_i^S)$, which does not depend upon k and, so, is determined by (33).

By substitution of (20) into the summed equation, the weights may be expressed in terms of the overt S-R transition probabilities through the N equations,

$$W_k^S = \frac{N \sum_i \left(\frac{P_{i(k)} - P^R}{P_{i(i)} - P^R}\right)^{1/2} \cdot \left(\frac{P_{k(i)} - P^R}{P_{k(k)} - P^R}\right)^{-1/2}}{\sum_h \sum_i \left(\frac{P_{i(h)} - P^R}{P_{i(i)} - P^R}\right)^{1/2} \cdot \left(\frac{P_{h(i)} - P^R}{P_{h(h)} - P^R}\right)^{-1/2}}. \tag{36}$$

Following a similar derivation, the response weights can be shown to be given by the N equations

$$W_k^R = \frac{N \sum_i \left(\frac{P_{(i)k} - P^S}{P_{(i)i} - P^S}\right)^{1/2} \cdot \left(\frac{P_{(k)i} - P^S}{P_{(k)k} - P^S}\right)^{-1/2}}{\sum_h \sum_i \left(\frac{P_{(i)h} - P^S}{P_{(i)i} - P^S}\right)^{1/2} \cdot \left(\frac{P_{(h)i} - P^S}{P_{(h)h} - P^S}\right)^{-1/2}}. \tag{37}$$

The Resolution of Psychological Distance into Orthogonal Coordinates in Psychological Space

In this section procedures will be set forth for estimating the distances D_{ik}^S and D_{ik}^R. In addition, a further reduction of the distances will be proposed so that the transition probability matrices can be reconstructed on the basis of still fewer quantities.

If, in (35), the same terms are used but both exponents are taken as positive, the weights may be eliminated to yield

$$\left(\frac{P_{ik}^S}{P_{ii}^S}\right)^{1/2} \cdot \left(\frac{P_{ki}^S}{P_{kk}^S}\right)^{1/2} = \exp(-D_{ik}^S). \tag{38}$$

In terms of the observed S-R transition probabilities, then, the distances are given by the N^2 equations

$$D_{ik}^S = -\log \left(\frac{P_{i(k)} - P^R}{P_{i(i)} - P^R}\right)^{1/2} \cdot \left(\frac{P_{k(i)} - P^R}{P_{k(k)} - P^R}\right)^{1/2}, \tag{39}$$

as may be seen by a substitution from (20).

Of the N^2 distances given by (39), $N(N - 1)/2$ have been supposed to vary independently. However, suppose the N stimulus points can be imbedded in a Euclidean space of K dimensions. In this case the distances can be reconstructed on the basis of just NK Cartesian coordinates, $X_{\alpha i}^S$, ($\alpha = 1, 2, \cdots, K$) and the generalized Pythagorean theorem

$$D_{ik}^S = \{\sum_\alpha (X_{\alpha i}^S - X_{\alpha k}^S)^2\}^{1/2}. \tag{40}$$

Thus an economy of description is possible in cases with $K < (N - 1)/2$.

Torgerson has presented a general method for determining a set of orthogonal coordinates, given a set of distances [25]. This procedure, as modified by Messick and Abelson [18] is as follows: Starting with the $N \times N$ symmetric matrix, $\mathbf{D}_{SS}$, of interstimulus distances, an $N \times N$ matrix, $\mathbf{B}_{SS}$, of scalar products of vectors from the centroid of the system of stimulus points to all pairs of points, S_i and S_k, is computed from

$$(41)\quad B^S_{ik} = \frac{1}{2N} \sum_j (D^S_{ij})^2 + \frac{1}{2N} \sum_j (D^S_{jk})^2 - \frac{1}{2} (D^S_{ik})^2 - \frac{1}{2N^2} \sum_g \sum_h (D^S_{gh})^2.$$

Young and Householder [27] have shown that, if this matrix is positive semidefinite (that is, if the points can be imbedded in a K-dimensional Euclidean space), it may be factored so that

$$(42)\qquad \mathbf{B}_{SS} = \mathbf{X}_{1S} \cdot \mathbf{X}'_{1S} + \mathbf{X}_{2S} \cdot \mathbf{X}'_{2S} + \cdots + \mathbf{X}_{KS} \cdot \mathbf{X}'_{KS},$$

where $\mathbf{X}_{\alpha S}$ represents a $N \times 1$ matrix (or column vector) giving the coordinates $\mathbf{X}_{\alpha i}$ for all stimuli S_i on the one dimension α, and where $\mathbf{X}'_{\alpha S}$ is the $1 \times N$ transpose of that matrix.

Now there are infinitely many possible factor decompositions of the form given in (42), each one corresponding to a different orientation of the orthogonal coordinate system in K-space. Which one of these is selected, however, can be a matter of arbitrary stipulation, since they all yield the same matrix $\mathbf{B}_{SS}$ and since the interstimulus distances, as given in (40), are invariant under orthogonal transformations of the coordinate system. In practice the individual dimensions (or factors) may be extracted in such a way that $\mathbf{X}_{1S}$ accounts for the largest possible variance among the original distances, $\mathbf{X}_{2S}$ for the largest possible variance in any direction orthogonal to the first dimension, and so on. In this way factoring may be terminated when the ability to reconstruct the original transition probability matrix is no longer significantly augmented by extraction of further dimensions. Various procedures for factoring $\mathbf{B}_{SS}$ in this way are available ([24], pp. 149–175, 473–510).

Exactly the same operations may be applied to the matrix of R-R transition probabilities, $\mathbf{P}_{RR}$, to obtain a symmetric matrix, $\mathbf{D}_{RR}$, of interresponse distances. The computation is implemented by the N^2 analogues of (39), namely,

$$(43)\qquad D^R_{ik} = -\log \left(\frac{P_{(i)k} - P^S}{P_{(i)i} - P^S}\right)^{1/2} \cdot \left(\frac{P_{(k)i} - P^S}{P_{(k)k} - P^S}\right)^{1/2}.$$

Once again, if the N response points can be imbedded in an L-dimensional Euclidean space, the distances may be reduced to NL orthogonal coordinates $X^R_{\beta i}$ $(\beta = 1, 2, \cdots, L)$ such that

$$D^R_{ik} = \{ \sum_\beta (X^R_{\beta i} - X^R_{\beta k})^2 \}^{1/2}. \tag{44}$$

The actual computation of the coordinates $X^R_{\beta i}$ will again be carried out by factoring a scalar product matrix $\mathbf{B}_{RR}$ so that

$$\mathbf{B}_{RR} = \mathbf{X}_{1R} \cdot \mathbf{X}'_{1R} + \mathbf{X}_{2R} \cdot \mathbf{X}'_{2R} + \cdots + \mathbf{X}_{LR} \cdot \mathbf{X}'_{LR} \, , \tag{45}$$

where $\mathbf{X}_{\beta R}$ is the column vector containing the coordinates for all responses on dimension β.

By substituting (40) and (44) into (30) and (31), (11) may now be stated in the more explicit form

$$\mathbf{P}_{SR} = \left[\frac{W^S_k \exp - \{ \sum_\alpha (X^S_{\alpha i} - X^S_{\alpha k})^2 \}^{1/2}}{\sum_h W^S_h \exp - \{ \sum_\alpha (X^S_{\alpha i} - X^S_{\alpha h})^2 \}^{1/2}} \right] \cdot \mathbf{J} \cdot \left[\frac{W^R_k \exp - \{ \sum_\beta (X^R_{\beta i} - X^R_{\beta k})^2 \}^{1/2}}{\sum_h W^R_h \exp - \{ \sum_\beta (X^R_{\beta i} - X^R_{\beta h})^2 \}^{1/2}} \right], \tag{46}$$

where the bracketed expressions represent matrices generated by allowing the indices i and k to run from 1 to N. Thus the complete set of N^2 *S-R* transition probabilities can be predicted on the basis of $2N$ weights, $(K + L)N$ coordinates, and the permutation matrix, $\mathbf{J}$, corresponding to the particular *S-R* assignment enforced.

Problems of Estimation

Certain practical difficulties arise in connection with the determination of the weights and coordinates owing to the fact that the probabilities, P_{ik}, are never known exactly but only estimated from the experimental data. The purpose of this section is to propose some approximate procedures which can be used when limitations on the number of subjects or the number of trials make these necessary.

First, with respect to the weights, the left-hand member of (35) will be extremely unstable, in a statistical sense, when P^S_{ik} and P^S_{ki} are small. A technique useful in alleviating this difficulty is the following: After the stimulus weights have been estimated in a preliminary way by (36), each row, i, of the matrix of quantities W^S_k/W^S_i given in (35) may be multiplied through by the tentative estimate for the corresponding W^S_i. In the resulting matrix, each of the N entries in column k will be an estimate of W^S_k. The final estimate may then be taken, for each column, as the median entry in that column. Exactly the same technique may be used to refine the response weight estimates.

In the estimation of distances the difficulty stems from the use of the logarithmic transformation of (39) and the consequent fact that, if $P_{i(k)}$

and $P_{k(i)}$ are small, a slight variation in these leads to a large variation in D_{ik}^{S}. The admission of such extreme instability in the determination of large interstimulus distances will have a disruptive influence on the factor solution used in obtaining the stimulus coordinates.

One might be inclined to redefine the distance between two widely separated stimuli, S_a and S_e, as the sum of distances over some connected path of smaller distances between them. Thus, in Fig. 2, it might be supposed, as an approximation, that

$$D_{ae}^{S} \cong D_{ab}^{S} + D_{bc}^{S} + D_{cd}^{S} + D_{de}^{S} .$$

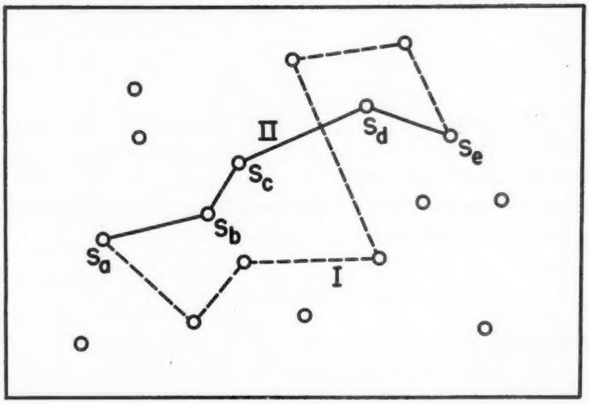

FIGURE 2

Different ways of approximating a large interstimulus distance by a sequence of smaller interstimulus distances. The circles represent stimuli in a two-dimensional psychological space.

The problem here is one of choosing between alternative paths such as I and II. However, such a selection should be guided by two general rules. First, a relatively direct path should be chosen. That is, the sum of distances over the path, $\sum D$, should be small. Second, a path should be chosen which does not contain any relatively large (and therefore unreliable) distances. That is, the largest distance in the path, $\max(D)$, should be small. Path I, then, is objectionable on both of these grounds.

Combining these rules, every distance in the matrix $\mathbf{D}_{SS}$ may be redefined as the sum, $\sum D$, of distances over that connected path for which the product

$$\max(D) \cdot \sum D \quad \text{is minimum.} \tag{47}$$

The small distances will generally remain unchanged during the re-estimation procedure. The large distances, however, will be subject to substantially less unsystematic error. The systematic error will presumably be somewhat augmented, however. The technique given here is not the statistically exact one which would have to take into account the number of events upon which each probability estimate is based.

Finally the application of the model is impeded insofar as it holds only for a given stage of learning, that is, over spans of trials for which the $P_{i(k)}$ remain relatively constant. The weights and distances could be estimated with greater reliability if they were based upon probabilities averaged over the entire learning session.

Now it is an implication of the model that the ratio of any two distances must be invariant over learning. Therefore some function, g, of time, t, exists such that

$$D^S_{ik(t)} = g(t) \cdot D^S_{ik} \, , \tag{48}$$

where $D^S_{ik(t)}$ is the psychological distance between S_i and S_k at some given time, t, as defined by (39), and where D^S_{ik} is some fixed distance between these stimuli which does not depend on t.

The fixed distances, D^S_{ik} , contain an arbitrary multiplicative constant which may be so adjusted that

$$\frac{1}{T} \int_0^T g(t) \, dt = 1. \tag{49}$$

It then follows that

$$D^S_{ik} = \frac{1}{T} \int_0^T D^S_{ik(t)} \, dt. \tag{50}$$

However, not only is this average computationally impractical, but it also heavily weights the large estimates for the $D^S_{ik(t)}$, which are based on small numbers of transitions and, so, are extremely unreliable.

Since the present model pertains to generalization at any given time rather than to the course of learning over time, no stipulation has been made regarding the function $g(t)$. However, the usual learning results indicate that the P_{ik} (for $i \neq k$) decline rapidly at first and then more slowly, from initial values of W^S_k/N towards their lower asymptotic bounds. This suggests that the average given in (50) may be more reliably approximated by the average

$$D^S_{ik} \cong -\log \left\{ \frac{1}{T} \int_0^T \exp \, [-D^S_{ik(t)}] \, dt - C^S \right\} \tag{51}$$

so long as the interstimulus distances do not cover too wide a range. This average, although it does discount the large, unstable distance estimates, requires the inclusion of a constant, C^S, in order to subtract out the essentially random transitions which account for the P_{ik} values of W^S_k/N before the subject has acquired any knowledge of the prevailing S-R assignment.

By transposing terms in (51), then, it may be seen that, for purposes of estimation, (30) is to be replaced by

$$P^S_{ik} \cong \frac{W^S_k[\exp(-D^S_{ik}) + C^S]}{\sum_h W^S_h[\exp(-D^S_{ih}) + C^S]}, \tag{52}$$

where P^S_{ik} applies to the entire learning session. Following through the derivation for (39), the psychological distances are found to be approximately given by

$$D^S_{ik} \cong -\log\left\{(1 + C^S)\left(\frac{P^S_{ik} \cdot P^S_{ki}}{P^S_{ii} \cdot P^S_{kk}}\right)^{1/2} - C^S\right\}. \tag{53}$$

Likewise, it is assumed that there exists a constant, C^R, such that

$$D^R_{ik} \cong -\log\left\{(1 + C^R)\left(\frac{P^R_{ik} \cdot P^R_{ki}}{P^R_{ii} \cdot P^R_{kk}}\right)^{1/2} - C^R\right\}. \tag{54}$$

In order to use (53), it is necessary to obtain an estimate for C^S. This may be done by calculating a set of stimulus coordinates under the assumption that $C^S = 0$. If, then, the quantities

$$\left(\frac{P_{i(k)} \cdot P_{k(i)}}{P_{i(i)} \cdot P_{k(k)}}\right)^{1/2}, \qquad (i, k = 1, 2, \cdots, N)$$

are plotted as a function of the distances reconstructed from (40), an asymptote, c, for large distances may be estimated by drawing a smooth curve through the data-points. If the responses have been selected so that $P^R = 0$, then $C^S = c/(1 - c)$. Exactly the same method can be used to estimate C^R, if $P^S = 0$.

The practical advantage of the counterbalancing technique mentioned in connection with (20) and (21) results from the following finding. In the application of (53) and (54), if P^R and P^S are small, they may be assumed equal to zero. The only appreciable consequence of this procedure appears to be a slight inflation of the estimates, respectively, for C^S and C^R.

With regard to the estimation of the stimulus and response weights, the constants C^S and C^R drop out in the derivation of (36) and (37). Therefore the weights can be approximately estimated from these equations, as they stand, even though the *S-R* transition probabilities are averaged over the entire learning session.

Appendix

Since, in all the experimental work to be reported $N = 9$, it will be useful to exhibit permutation matrices, $\mathbf{J}_m$, having the property that over all subjects, every pair of responses is assigned to each pair of stimuli the same number of times. In order to do this, it is convenient to introduce $\mathbf{0}_3$, the

3×3 null matrix; $\mathbf{I}_3$, the 3×3 identity matrix; $\mathbf{H}_3$, given by

$$\mathbf{H}_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix};$$

and $\mathbf{K}_3^{ii}$, the 3×3 matrix with elements $K_{gh} = 1$, if $g = i$ and $h = j$, and $K_{gh} = 0$, otherwise. If for positive integers r,

$$\mathbf{J}_{[3r-1]} = \mathbf{J}_{[3r]} = \begin{bmatrix} \mathbf{H}_3 & \mathbf{O}_3 & \mathbf{O}_3 \\ \mathbf{O}_3 & \mathbf{H}_3 & \mathbf{O}_3 \\ \mathbf{O}_3 & \mathbf{O}_3 & \mathbf{H}_3 \end{bmatrix},$$

$$\mathbf{J}_{[9r-5]} = \mathbf{J}_{[9r-2]} = \begin{bmatrix} \mathbf{O}_3 & \mathbf{I}_3 & \mathbf{O}_3 \\ \mathbf{O}_3 & \mathbf{O}_3 & \mathbf{I}_3 \\ \mathbf{I}_3 & \mathbf{O}_3 & \mathbf{O}_3 \end{bmatrix} \cdot \mathbf{J}_{[3r-1]},$$

$$\mathbf{J}_{[9r+1]} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{O}_3 & \mathbf{O}_3 \\ \mathbf{O}_3 & \mathbf{H}_3 & \mathbf{O}_3 \\ \mathbf{O}_3 & \mathbf{O}_3 & \mathbf{H}_3' \end{bmatrix} \cdot \begin{bmatrix} \mathbf{K}_3^{11} & \mathbf{K}_3^{21} & \mathbf{K}_3^{31} \\ \mathbf{K}_3^{12} & \mathbf{K}_3^{22} & \mathbf{K}_3^{32} \\ \mathbf{K}_3^{13} & \mathbf{K}_3^{23} & \mathbf{K}_3^{33} \end{bmatrix} \cdot \mathbf{J}_{[9r-5]},$$

then the permutation matrix for any subject m (starting with some arbitrary initial assignment, $\mathbf{J}_{[1]}$) is given by

$$\mathbf{J}_m = \mathbf{J}_{[m]} \cdot \mathbf{J}_{[m-1]} \cdot \cdots \cdot \mathbf{J}_{[1]} .$$

The first 36 matrices formed in this way assign every pair of responses to each pair of stimuli just once. The second 36 assign the same pairs of responses to the same pairs of stimuli with the orders reversed from those of the first 36 assignments. Thereafter the same assignments are repeated with every succeeding 72 matrices. Thus, if $N = 9$, M can be any multiple of 72. Indeed, since P_{ik}^R will generally not be far from P_{ki}^R, a satisfactory degree of counterbalancing can probably be obtained with only the first 36 assignments.

REFERENCES

[1] Attneave, F. Dimensions of similarity. *Amer. J. Psychol.*, 1950, **63**, 516-556.

[2] Blumenthal, L. M. Theory and applications of distance geometry. Oxford: Clarendon Press, 1953.

[3] Brown, J. S., Bilodeau, E. A., and Baron, M. R. Bidirectional gradients in the strength of a generalized voluntary response to stimuli on a visual-spatial dimension. *J. exp. Psychol.*, 1951, **41**, 52-61.

[4] Busemann, H. The geometry of geodesics. New York: Academic Press, 1955.

[5] Bush, R. R. and Mosteller, F. A model for stimulus generalization and discrimination. *Psychol. Rev.*, 1951, **58**, 413-423.

[6] Bush, R. R. and Mosteller, F. Stochastic models for learning. New York: Wiley, 1955.

[7] Duncan, C. P. Development of response generalization gradients. *J. exp. Psychol.*, 1955, **50**, 26-30.

[8] Estes, W. K. Towards a statistical theory of learning. *Psychol. Rev.*, 1950, **57**, 94-107.

[9] Frick, F. C. An analysis of an operant discrimination. *J. Psychol.*, 1948, **26**, 93-123.

[10] Gibson, E. J. Sensory generalization with voluntary reactions. *J. exp. Psychol.*, 1939, **24**, 237-253.

[11] Gulliksen, H. and Wolfle, D.L. A theory of learning and transfer: I. *Psychometrika*, 1938, **3**, 127-149.

[12] Guttman, N. and Kalish, H. I. Discriminability and stimulus generalization. *J. exp. Psychol.*, 1956, **51**, 79-88.

[13] Hovland, C. I. The generalization of conditioned responses: I. The sensory generalization of conditioned responses with varying frequencies of tone. *J. gen. Psychol.*, 1937, **17**, 125-148.

[14] Hovland, C. I. Human learning and retention. In S. S. Stevens (Ed.), Handbook of experimental psychology. New York: Wiley, 1951.

[15] Hull, C. L. Principles of behavior. New York: Appleton-Century, 1943.

[16] Kelley, J. L. General topology. New York: Van Nostrand, 1955.

[17] Messick, S. J. Some recent theoretical developments in multidimensional scaling. *Educ. psychol. Measmt*, 1956, **16**, 82-100.

[18] Messick, S. J. and Abelson, R. P. The additive constant problem in multidimensional scaling. *Psychometrika*, 1956, **12**, 1-15.

[19] Margolius, G. Stimulus generalization of an instrumental response as a function of the number of reinforced trials. *J. exp. Psychol.*, 1955, **49**, 105-111.

[20] Noble, M. E. and Bahrick, H. P. Response generalization as a function of intratask response similarity. *J. exp. Psychol.*, 1956, **51**, 405-412.

[21] Pillsbury, W. B. A study in apperception. *Amer. J. Psychol.*, 1897, **8**, 315-393.

[22] Plotkin, L. Stimulus generalization in Morse code learning. *Arch. Psychol.*, 1943, **40**, No. 287.

[23] Rosenbaum, G. Stimulus generalization as a function of level of experimentally induced anxiety. *J. exp. Psychol.*, 1953, **45**, 35-43.

[24] Thurstone, L. L. Multiple-factor analysis. Chicago: Univ. Chicago Press, 1947.

[25] Torgerson, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika*, 1952, **17**, 401-420.

[26] Woodworth, R. S. and Schlosberg, H. Experimental psychology. New York: Holt, 1955.

[27] Young, G. and Householder, A. S. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 1938, **3**, 19-22. ✓

Manuscript received 8/17/56

Revised manuscript received 4/29/57

PSYCHOMETRIKA—VOL. 22, NO. 4
DECEMBER, 1957

THE RELATIONSHIP BETWEEN FACTORIAL COMPOSITION OF TEST ITEMS AND MEASURES OF TEST RELIABILITY*

JOHN W. COTTON
DONALD T. CAMPBELL
AND
R. DANIEL MALONE
NORTHWESTERN UNIVERSITY

For continuous distributions associated with dichotomous item scores, the proportion of common-factor variance in the test, H^2, may be expressed as a function of intercorrelations among items. H^2 is somewhat larger than the coefficient α except when the items have only one common factor and its loadings are restricted in value. The dichotomous item scores themselves are shown not to have a factor structure, precluding direct interpretation of the Kuder-Richardson coefficient, $r_{K\text{-}R}$, in terms of factorial properties. The value of $r_{K\text{-}R}$ is equal to that of a coefficient of equivalence, $H^2{}_\Phi$, when the mean item variance associated with common factors equals the mean inter-item covariance. An empirical study with synthetic test data from populations of varying factorial structure showed that the four parameters mentioned may be adequately estimated from dichotomous data.

Factor structure and test reliability, r_{tt} , are closely connected in testing theory. In Cronbach's [2] joint treatment of these two topics he distinguishes four kinds of reliability coefficients: (1) stability, (2) stability-and-equivalence, (3) equivalence, and (4) hypothetical-self-correlation. Each, he says, is uniquely characterized by the particular factor score variances assigned to error variance, σ_e^2 . Thus the coefficient of equivalence is defined by a general formula, $r_{tt} = 1 - (\sigma_e^2/\sigma_t^2)$, where σ_t^2 is the variance of total test scores and σ_e^2 includes both the variance of specific factors for each item and the residual error variance. Stated differently, the coefficient of equivalence tells "the degree to which the test score indicates the status of the individual at the present instant in the general and group factors defined by the test" [2]. A definition of another coefficient, such as that of stability, would employ the same general formula with a new specification of error variance.

In a later analysis of the Kuder-Richardson ([7], formula 20) coefficient of equivalence, r_{K-R}, Cronbach [3] has stated in greater detail how total

*This study was supported in part by an Air Force project (Contract Number AF18(600-170), monitored by the Crew Research Laboratory, Air Force Personnel and Training Research Center, Randolph Air Force Base, Randolph Field, Texas. Permission is granted for reproduction, translation, publication, use and disposal in whole and in part by or for the United States Government. Further support was given by the Northwestern University Graduate School. The computational assistance of Mr. Norman Miller is acknowledged. Professor Meyer Dwass provided mathematical advice both directly and indirectly relevant to the paper.

test variance depends upon the factor loadings of the individual items. He repeats his earlier argument that a general coefficient, α, of which r_{K-R} is a special case, is the proportion of test variance due to common factors when a special assumption holds: the mean common-factor variance within items must equal the mean interitem covariance. This conclusion is weakened by the recognition that the assumption does not hold for an interitem correlation matrix with rank greater than one. Thus r_{K-R} is the proportion of common-factor variance only when there is but one common factor. Cronbach presumes that α will closely estimate this proportion even with multifactored cases unless the test contains distinct clusters.

The present paper argues that a strict interpretation of the factorial hypothesis requires a reanalysis of Cronbach's notions. This reanalysis requires a separate treatment of factorial structure and of reliability whenever dichotomously scored items comprise the test under analysis. In brief, a factor structure of such items and the related factor structure of the total test exist for continuous distributions underlying each dichotomous distribution. These structures do not exist for the dichotomous distributions. On the other hand, since the scoring of items and of the total test employs the dichotomized scores, reliability measures must be properties of the dichotomous distribution.

A consequence of this reasoning will be the specification of two distinct α coefficients: α for the continuous case, and α_Φ or r_{K-R} for the dichotomous case. The conditions under which α will equal the proportion of common-factor variance in the total test variance, H^2, will be examined, extending Cronbach's statement on the single-factoredness requirement. The coefficient α_Φ or r_{K-R} will be shown to approximate a coefficient of equivalence, H_Φ^2 , for a test with dichotomously scored items. Thus, contrary to Cronbach's belief, r_{K-R} will *never* estimate common-factoredness, even for single-factored tests, and will estimate a coefficient of equivalence only under special conditions even more restrictive than single-factoredness.

Theory of the Continuous Case

Thurstone's multiple-factor theory has as its basis the hypothesis ([8], pp. 69–74):

$$s_{ip} = \sum_{m=1}^{r} a_{im}x_{mp} + b_iy_{ip} + e_i\,\eta_{ip}\ , \tag{1}$$

where

s_{ip} ≡ standard score for person p on item i of a test ($i = 1, 2, \cdots, n$),
a_{im} ≡ loading of common factor m on item i ($m = 1, 2, \cdots, r$),
x_{mp} ≡ standard score for person p on common factor m,
b_i ≡ loading of item i on a factor specific to item i,

$$e_i = \sqrt{1 - \sum_{m=1}^{r} a_{im}^2 - b_i^2} \equiv \text{loading of error on item } i,$$

and y_{ip} and η_{ip} have definitions analogous to that for x_{mp} . The x_{mp} , y_{ip} , and η_{ip} are independent of each other. The values of a_{im} , b_i , and e_i are parameters of the test.

The item raw scores, $X_{ip} = \sigma_i s_{ip} + \mu_i$, where μ_i and σ_i are the population mean and standard deviation of the x_{ip} , may be summed to give total test scores

$$(2) \quad T_p = \sum_{i=1}^{n} X_{ip} = \sum_{m=1}^{r} \left(\sum_{i=1}^{n} \sigma_i a_{im} x_{mp} \right) + \sum_{i=1}^{n} \sigma_i b_i y_{ip} + \sum_{i=1}^{n} \sigma_i e_i \eta_{ip} + \sum_{i=1}^{n} \mu_i \, .$$

On a second administration of the test, the standard scores and total test scores will be given by

$$(3) \qquad s'_{ip} = \sum_{m=1}^{r} a_{im} x_{mp} + b_i y_{ip} + e_i \eta'_{ip} \, ,$$

and

$$(4) \qquad T'_p = \sum_{m=1}^{r} \left(\sum_{i=1}^{n} \sigma_i a_{im} x_{mp} \right) + \sum_{i=1}^{n} \sigma_i b_i y_{ip} + \sum_{i=1}^{n} \sigma_i e_i \eta'_{ip} + \sum_{i=1}^{n} \mu_i \, .$$

Equations (3) and (4) differ from (1) and (2) only in that η_{ip} is a random variable changing in value to η'_{ip} on the second administration of the test. The η'_{ip} are independent of all previous variables.

Following Wilks ([10], pp. 33–35) and remembering that the variances of the x_{mp} , y_{ip} , η_{ip} and η'_{ip} are all unity lead to

$$(5) \qquad \text{var}\,(T) = \text{var}\,(T') = \sum_{m=1}^{r} \left(\sum_{i=1}^{n} \sigma_i a_{im} \right)^2 + \sum_{i=1}^{n} \sigma_i^2 b_i^2 + \sum_{i=1}^{n} \sigma_i^2 e_i^2 \, .$$

Equation (5) is the same in substance as Cronbach's [2] equations (2) and (3). Similarly the covariance of total test scores on successive administrations is given by

$$(6) \qquad \text{cov}\,(T, T') = \sum_{m=1}^{r} \left(\sum_{i=1}^{n} \sigma_i a_{im} \right)^2 + \sum_{i=1}^{n} \sigma_i^2 b_i^2 \, .$$

The ratio of cov (T, T') to var (T) is the hypothetical self-correlation $r_{TT'}$ of the continuous scores T_p and T'_p . It is also seen from (5) and (6) to be the ratio of variance contributed by common and specific factors to total test variance.

The relative contribution of each separate common or specific factor to var (T) is

$$(7) \qquad A_m^2 = \left(\sum_{i=1}^{n} \sigma_i a_{im} \right)^2 \Big/ \text{var}\,(T);$$

$$B_i^2 = \sigma_i^2 b_i^2 / \text{var}\ (T). \tag{8}$$

A_m^2 and B_i^2 may also be called the squared factor loadings of common factor m for the total test and of specific factor i for that test.

The relative contribution of common factors to var (T) has previously been termed H^2:

$$H^2 = \sum_{m=1}^{r} A_m^2 = \sum_{m=1}^{n} \left(\sum_{i=1}^{n} \sigma_i a_{im} \right)^2 \Big/ \text{var}\ (T). \tag{9}$$

H^2 may also be considered a coefficient of equivalence for the continuous case since $1 - H^2$ includes both specific variance and residual variance.

It may be of note to remark that with a centroid solution having only positive first factor loadings, Thurstone has shown $\sum_{i=1}^{n} a_{im} = 0$ for $m \neq 1$. This implies that $A_m^2 = 0$ for $m \neq 1$ and, consequently, that $H^2 = A_1^2$ centroid.

Equation (9) must be revised to define H^2 in terms of parameters more readily determined than the a_{im}'s. By definition, r_{ij}, the correlation coefficient between the item i and item j in a single test administration, is the expected value of $(s_{ip}s_{pj})$. This definition coupled with (1) yields

$$r_{ij} = \sum_{m=1}^{r} a_{im} a_{jm}\ , \qquad (i \neq j); \tag{10}$$

$$r_{ij} = \sum_{m=1}^{r} a_{im} a_{jm} + b_i^2 + e_i^2 = 1, \qquad (i = j). \tag{11}$$

The expressions (10) and (11) in turn imply that (5) may be rewritten

$$\text{var}\ (T) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_i \sigma_j r_{ij}\ . \tag{12}$$

Equation (12) will be used to restate the denominator of (9). Reanalyzing the numerator of that equation, define r_{ij}^* as the correlation coefficient between the items i and j on successive administrations of the test, excluding the contribution of specific factors. In this case

$$r_{ij}^* = r_{ij} = \sum_{m=1}^{r} a_{im} a_{jm}\ , \qquad (i \neq j), \tag{13}$$

and

$$r_{ij}^* = h_i^2 = \sum_{m=1}^{n} a_{im}^2\ , \qquad (i = j), \tag{14}$$

where h_i^2 is the proportion of common-factor variance to total variance for item i. Equations (13) and (14) lead to the conclusion

$$\sum_{m=1}^{r} \left(\sum_{i=1}^{n} \sigma_i a_{im} \right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_i \sigma_j r_{ij}^*\ . \tag{15}$$

Equations (12) and (15) may now be employed to rewrite (9) as

$$H^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j r_{ij}^{*}\Big/\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j r_{ij}\ . \tag{16}$$

This equation is directly applicable to situations in which the σ_i , σ_j , and r_{ij} are known but the factor loadings themselves have not been determined.

The coefficient α for the continuous case may now be compared with H^2. Cronbach's equation (24) ([3], p. 305) in our notation is

$$\alpha = \frac{n}{n-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\sigma_i\sigma_j r_{ij}\Big/\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j r_{ij}\ . \tag{17}$$

For α to equal H^2, the proportion of common-factor variance to total test variance, the numerators of (16) and (17) must be equal. This requirement reduces by means of (13) and (14) to

$$\sum_{i=1}^{n}\sigma_i^2h_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j r_{ij}^{*} = \frac{n}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j r_{ij}^{*}\ , \qquad (i \neq j). \tag{18}$$

Simple algebraic manipulations lead to an equivalent condition:

$$\sum_{i=1}^{n}\sigma_i^2h_i^2 - \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j r_{ij}^{*}\right)\Big/n = 0. \tag{19}$$

It should be noted that the case $i = j$ is not excluded in equation (19) as it was in (18).

Finally by (14) and (15), (19) leads to

$$\sum_{m=1}^{r}\left[\sum_{i=1}^{n}\sigma_i^2a_{im}^2 - \left(\sum_{i=1}^{n}\sigma_i a_{im}\right)^2\Big/n\right] = 0. \tag{20}$$

But this is equivalent to the assertion that the standard deviation of the $\sigma_i a_{im}$ must be zero for every m, or equivalent to the assertion that $a_{im} = k_m/\sigma_i$, where k_m is constant over i but may vary over m. Combining this result with (13) and (14) gives a general term for the reduced correlation matrix:

$$r_{ij}^{*} = \sum_{m=1}^{r}k_m^2\Big/\sigma_i\sigma_j\ .$$

That matrix would then have rank 1, and a single common-factor solution with loadings equal to

$$\sqrt{\sum_{m=1}^{r}k_m^2}\Big/\sigma_i\ .$$

In summary of this point, Cronbach's requirement of a single common factor for equality of H^2 and α may be expanded to require that each item have a single common-factor loading inversely proportional to its σ_i . Although this condition is both necessary and sufficient, H^2 and α may be *almost* equal without satisfaction of the condition. For estimation procedures employing dichotomous data, the restrictive assumption $\sigma_i = \sigma$ for all i will become necessary. In that case σ will replace σ_i and σ_j where they occur in (2) through (20). In particular H^2 will be defined in a manner almost equivalent to Jackson and Ferguson's equation for $r_{TT'}$ ([6], eq. 67)

$$H^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij}^* \Big/ \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij} \tag{21}$$

in place of that given in (16), and α will be given by

$$\alpha = \frac{n}{n-1} \sum_{\substack{i=1 \\ (i \neq j)}}^{n} \sum_{j=1}^{n} r_{ij} \Big/ \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij} \,. \tag{22}$$

Similarly the condition (20) required for equivalence of H^2 and α will reduce to a requirement of a single common factor with a *constant* loading equal to

$$\sqrt{\sum_{m=1}^{r} (k_m^2/\sigma^2)} = \sqrt{\sum_{m=1}^{r} a_{im}^2}\cdot$$

Theory of the Dichotomous Case

Suppose that a population proportion P_i of persons pass item i of the n-item test previously discussed. Can we describe the standard scores corresponding to 0 (fail) and 1 (pass) scores by an expression having the form of (1)? All persons failing the item will have $s_{ip} = -P_i/\sqrt{P_i(1 - P_i)}$; all persons passing the item will have $s_{ip} = (1 - P_i)/\sqrt{P_i(1 - P_i)}$. Now any nontrivial application of (1) implies the existence of at least one nonzero a_{im}, two distinct values of x_{mp}, a nonzero e_i, and two distinct values of η_{ip}. Consequently (1) requires that there be at least three values of s_{ip} in the simplest case, occurring when $a_{im} = \sqrt{2}/2 = e_i$, $x_{mp} = \pm 1$, and $\eta_{ip} = \pm 1$. Since the x_{mp} and η_{ip} are independent, s_{ip} will take on one of three values, for various persons, $\sqrt{2}$, 0, and $-\sqrt{2}$ rather than the s_{ip} values given above.

But this simplest case completely refutes the proposition that dichotomous item scores have a factorial structure of their own. The factorial hypothesis of (1) will always imply the existence of three or more different score values on each item. One should never, then, suppose that dichotomous distributions may be generated by (1). Correspondingly, one should never factor analyze a product moment correlation matrix based on 0 and 1 scores (i.e., a matrix of phi coefficients). Although this matrix may be expressed as a product of a "factor" matrix by its transpose, that "factor" matrix will have no direct application to anything in factor theory. The misbehavior of factor loadings based on phi coefficients has been previously observed [5, 9]. The theoretical basis for this misbehavior has not been treated in the present manner.

In denying the factorial interpretability of dichotomized scores as basic data, we do not completely foreswear an interest in the dichotomous case. As a later section will show, dichotomous data can be used in such a way as to estimate parameters, such as H^2 and α, which are characteristic of the underlying continuous distributions.

We now turn to the matter of reliability determinations for total test scores obtained by summing pass-fail item scores. The hypothetical self-correlation, $r_{TT'_\Phi}$, may be defined as

$$(23) \qquad r_{TT'_\Phi} = \text{cov}_\Phi\,(T, T')/\text{var}_\Phi\,(T) = \sum_{i=1}^{n} \sum_{j=1}^{n} \text{cov}\ \Phi_{ij} \Big/ \text{var}\ \Phi\,(T),$$

where the Φ subscripts are employed to indicate that dichotomous item scores are employed, causing all interitem correlations to be phi coefficients, Φ_{ij}.

Use of the facts that

$$\sigma_{\Phi_i} = \sqrt{P_i(1 - P_i)} \quad \text{and} \quad \sigma_{\Phi_j} = \sqrt{P_j(1 - P_j)}\,,$$

helps (23) become

$$(24) \qquad r_{TT'_\Phi} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{P_i(1 - P_i)P_j(1 - P_j)}\,\Phi_{ij} \Big/ \text{var}_\Phi\,(T).$$

This hypothetical self-correlation may be called a phi coefficient analogue of $r_{TT'}$. Its value will be less than $r_{TT'}$ and, unlike $r_{TT'}$, it will be sensitive to changes in P_i from item to item.

To obtain a coefficient of equivalence, H_Φ^2, for the dichotomous case, any quantity attributable to specific factors in the underlying item scores will be excluded from the numerator of (24). This is done by defining H_Φ^2 as

$$(25) \qquad H_\Phi^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{P_i(1 - P_i)P_j(1 - P_j)}\,\Phi_{ij}^* \Big/ \text{var}_\Phi\,(T),$$

where $\Phi_{ij}^* = \Phi_{ij}$ when $i \neq j$, and $\Phi_{ij}^* = \Phi(h_i^2)$ when $i = j$. $\Phi(h_i^2)$ is simply the phi coefficient obtained by entering the Chesire, Saffir, and Thurstone [1] tetrachoric correlation charts backwards, using $r_{ij} = h_i^2$ and P_i to obtain the fourfold table necessary for computation of the phi coefficient associated with that h_i^2. H_Φ^2 is not a measure of common factoredness, but factor structure plus the P_i and P_j have complete control over it. It is the idealized test-retest correlation for total test scores based on pass-fail data when the self-correlation of items has been reduced to exclude specific factor contribution to the underlying distribution. H_Φ^2 may also be called the correlation between tests which are matched item by item for common-factor structure.

An analogue of α, α_Φ or r_{K-R}, has already been mentioned:

$$(26) \quad \alpha_\Phi = r_{K\text{-}R} = \frac{n}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{P_i(1 - P_i)P_j(1 - P_j)}\,\Phi_{ij} \Big/ \text{var}_\Phi\,(T).$$

A comparison of (25) and (26) shows that r_{K-R} is a coefficient of equivalence when

$$(27) \qquad \overline{P_i(1 - P_i)\Phi^*_{ii}} = \overline{\sqrt{P_i(1 - P_i)P_j(1 - P_j)}\,\Phi_{ij}}\ , \qquad (i \neq j).$$

This condition is analogous to requirement (18), and both are equivalent for the case of parallel tests to Jackson and Ferguson's [6] assumption that the mean interitem covariance between tests be equal to the mean interitem covariance within tests.

In the special case where $P_i(1 - P_i)$ is constant over all i and σ_i is also constant over all i, the one-to-one correspondence between r_{ij} and Φ_{ij} will imply that the *modified* condition (20) following equation (22) is necessary and sufficient for r_{K-R} to be a coefficient of equivalence.

The Estimation of H^2, H^2_Φ, α, and r_{K-R}

A crude estimation procedure for the four parameters, H^2, H^2_Φ, α, and r_{K-R}, is to replace all individual components of (16), (25), (17), and (26) by their estimators. Special problems arising in the application of (16) and (17) when only dichotomous data are available are insoluble until some assumption about the σ_i and σ_j values for the underlying continuous distribution is made.

Obviously the P_i and P_j values of the multivariate dichotomous population may be assigned independently of the σ_i and σ_j values for the underlying continuous population. Then the pass-fail splits defining the Φ_{ij} will be independent of the σ_i and σ_j, depending only on the P_i, P_j, and r_{ij}. Thus neither sample values nor parameter values from the dichotomous population alone will give any information about the σ_i and σ_j.

Any arbitrary assumption about the σ_i and σ_j would permit estimation of H^2 and α with (16) and (17). In general the assumption $\sigma_i = \sigma$ for all i leading to (21) and (22) seems most satisfactory. Therefore, it will be employed throughout the remainder of this paper.

The estimation of H^2 and α from dichotomous data may be performed for multivariate normally distributed underlying item scores, at least, by replacing each population r_{ij} in (21) and (22) by a tetrachoric correlation coefficient obtained from the Chesire, Saffir, and Thurstone tables. Then h^2_i is estimated as the highest tetrachoric coefficient of column i in the tetrachoric matrix. Improved estimation procedures would no doubt result from an attempt to obtain maximum likelihood estimates for H^2, H^2_Φ, α, and r_{K-R}.

One may well ask the justification of estimating H^2 and α when sample item scores are all dichotomous and the alleged continuous underlying distributions seem but a convenient fiction. Three answers may be given to this. (*a*) There is no other way to make a statement about the proportion of common-factoredness in such tests. (*b*) The coefficients H^2 and α are upper bounds on H^2_Φ and r_{K-R}, showing the degree of improvement in the test which could be obtained by continuous rather than dichotomous scoring of each item. With such materials as tests of physical strength, it may be quite feasible to increase substantially the coefficient of equivalence by a change in scoring methods without changing the test items employed. (*c*) The sensitivity of r_{K-R} to variation in P_i is well known. Many attempts to control test homogeneity have centered on control of the P_i . Comparisons between H^2 and H_Φ^2 and between α and r_{K-R} will serve to emphasize that a test with high r_{K-R} because of homogeneous P_i may nevertheless have less factorial homogeneity than a test with a lower r_{K-R} .

A Sampling Study

The unit of data for this study is a sextuplet of numbers representing item scores on a six-item "test" given to a hypothetical "subject." Eight samples of 500 such subjects' item scores were obtained, one sample from each of eight populations defined by specifying the factor loadings of the items in the test associated with that population of scores. The factor loadings selected are presented in Table 1.

TABLE 1

Item Factor Loadings and Test Parameter Values for Eight Contrived Populations of Scores from Six-Item Tests

		Loading for Item Number					
Population		1	2	3	4	5	6
I	$\underline{a_{i1}}$	.7746	.7746	.7746	.7746	.7746	.7746
	$\underline{a_{i1}}$	.8660	.8660	.8660	0	0	0
II	$\underline{a_{i2}}$	0	0	0	.8660	.8660	.8660
	$\underline{a_{i1}}$	.9045	.9045	0	0	0	0
III	$\underline{a_{i2}}$	0	0	.9045	.9045	0	0
	$\underline{a_{i3}}$	0	0	0	0	.9045	.9045
	$\underline{a_{i1}}$	.5884	.5884	0	.5884	.5884	0
IV	$\underline{a_{i2}}$	0	.5884	.5884	0	.5884	.5884
	$\underline{a_{i3}}$	.5884	0	.5884	.5884	0	.5884
	$\underline{a_{i1}}$	.8321	.8321	0	.8321	.8321	.0
V	$\underline{a_{i2}}$	0	.4161	.4161	0	.4161	.4161
	a_{i3}	.4161	0	.4161	.4161	0	.4161
VI	$\underline{a_{i1}}$	.5291	.5291	.5291	.5291	.5291	.5291
VII	$\underline{a_{i1}}$	.3780	.3780	.3780	.3780	.3780	.3780
	$\underline{a_{i1}}$	.4472	.4472	0	.4472	.4472	0
VIII	$\underline{a_{i2}}$	0	.2236	.2236	0	.2236	.2236
	$\underline{a_{i3}}$	.2236	0	.2236	.2236	0	.2236

Given the a_{im} and e_i values for any population, we employed a table of random normal numbers ([4], Table 2) with $\mu = 0$, $\sigma = 1$, to obtain a different set of 500 x_{mp} or x_{ip} values for each factor and for error. Equation (1) was employed to obtain s_{ip} values for each person on each item.

After all s_{ip} values had been obtained for a test, the 500 scores for each item were dichotomized on the basis of a pass-fail criterion. This permitted determination of tetrachoric correlation coefficients for use in computing $\tilde{H}^2$ and $\tilde{\alpha}$, estimators of H^2 and α. The dichotomized scores were also employed in calculating $\tilde{H}^2_\Phi$ and $\tilde{r}_{K-R}$. In the determination of H^2_Φ and r_{K-R}, the sample proportions were used in place of population P_i, introducing some inaccuracy in their values.

The dichotomization of item scores was performed twice, once with the sample proportion of 1 scores, p_i, fixed at .50 for every item in every test and once with p_i variable from item to item. In the latter case p_i ranged from .25 to .75, with $p_1 = .25$, $p_2 = .35$, $\cdots$, $p_6 = .75$ for every test. Each of the dichotomizations led to a distinct set of estimates of parameter values.

Results

Table 2 presents a comparison of sample and population values of all four coefficients, both for the fixed p_i case and the variable p_i case. Within the limits of the samples employed, our estimators are quite satisfactory. The mean constant error of any estimator across eight populations never exceeds .002 in absolute values, and no individual coefficient is in error by more than .058.

TABLE 2

A Comparison of Parameter and Estimator Values for a Six-Item Synthetic Test

Population			Fixed p_i						Variable p_i					
	H^2	α	$\tilde{H}^2$	$\tilde{\alpha}$	H^2_Φ	$\tilde{H}^2_\Phi$	r_{K-R}	$\tilde{r}_{K-R}$	$\tilde{H}^2$	$\tilde{\alpha}$	H^2_Φ	$\tilde{H}^2_\Phi$	r_{K-R}	$\tilde{r}_{K-R}$
I	.900	.900	.929	.914	.800	.840	.800	.822	.929	.905	.788	.820	.777	.775
II	.900	.720	.885	.734	.779	.789	.623	.623	.904	.689	.749	.768	.555	.565
III	.900	.540	.893	.579	.756	.748	.454	.472	.897	.504	.751	.740	.444	.403
IV	.900	.810	.911	.838	.779	.801	.694	.726	.890	.823	.766	.725	.661	.673
V	.900	.810	.882	.801	.804	.783	.713	.702	.895	.813	.787	.792	.668	.676
VI	.700	.700	.702	.680	.575	.575	.575	.549	.714	.670	.549	.540	.544	.527
VII	.500	.500	.558	.521	.389	.431	.389	.398	.587	.542	.362	.410	.360	.356
VIII	.500	.450	.488	.416	.372	.358	.336	.304	.416	.328	.356	.309	.316	.251

In single-factored populations I, VI, and VII, H^2 and α are equal, and H_Φ^2 and r_{K-R} are equal for the homogeneous p_i case only. In some of the other populations these coefficients differ markedly, indicating that α and r_{K-R} are poor approximations to coefficients of equivalence in those cases. A large discrepancy between sample values, $\tilde{H}^2$ and $\tilde{\alpha}$, or $\tilde{H}_\Phi^2$ and $\tilde{r}_{K-R}$, appears indicative of multiple-factoredness or of wide dispersion of a single common factor's loadings. The converse statement is not true.

Unlike H^2 and α, which are invariant under different dichotomizations of the same underlying scores, H_Φ^2 and r_{K-R} are reduced by introducing heterogeneity of the p_i . The estimators $\tilde{H}_\Phi^2$ and $\tilde{r}_{K-R}$ also show this effect.

REFERENCES

[1] Chesire, L., Saffir, M., and Thurstone, L. L. Computing diagrams for the tetrachoric correlation coefficient. Chicago: Univ. of Chicago Bookstore, 1933.

[2] Cronbach, L. J. Test "reliability": its meaning and determination. *Psychometrika*, 1947, 12, 1-16.

[3] Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.

[4] Dixon, W. J. and Massey, F. J., Jr. Introduction to statistical analysis. New York: McGraw-Hill, 1951.

[5] Ferguson, G. A. The factorial interpretation of test difficulty. *Psychometrika*, 1941, 6, 323-329.

[6] Jackson, R. W. B. and Ferguson, G. A. Studies on the reliability of tests. Toronto: Department of Educational Research, Bulletin No. 12, 1941.

[7] Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-166.

[8] Thurstone, L. L. Multiple-factor analysis. Chicago: Univ. of Chicago Press, 1947.

[9] Wherry, R. J. and Gaylord, R. H. The concept of test and item reliability in relation to factor-pattern. *Psychometrika*, 1943, 8, 247-269.

[10] Wilks, S. S. Mathematical statistics. Princeton: Princeton Univ. Press, 1943.

Manuscript received 1/3/56

Revised manuscript received 4/24/57

PSYCHOMETRIKA—VOL. 22, NO. 4
DECEMBER, 1957

THE USE OF CONFIGURAL ANALYSIS FOR THE EVALUATION OF TEST SCORING METHODS*

H. G. OSBURN

SOUTHERN ILLINOIS UNIVERSITY

AND

ARDIE LUBIN

WALTER REED ARMY INSTITUTE OF RESEARCH

A method based on configural analysis has been given whereby test scoring techniques can be evaluated to see if they have optimal validity. Configural analysis has also been used to show how three well known item scoring techniques, multiple regression, total score, and multiple cut-off, imply (for optimal validity) certain conditions on the answer pattern means. The method is illustrated by a worked example.

The purpose of this article is to demonstrate a method whereby test scoring techniques can be evaluated to see if they have maximum validity. In a previous paper [3] a technique of pattern scoring of test items for the prediction of a quantitative criterion was presented. The basic notion used was that of a configural scale, defined as follows: given a test of t items and a quantitative criterion, form all possible answer patterns and assign to each subject a score which is the mean criterion score for all subjects in his answer pattern. This set of scores is called a configural scale. It was shown that, in the analysis sample, of all possible ways of scoring the t items, the configural scale provides the best least squares prediction of the criterion. It was further shown that the configural scale could be represented exactly by a polynomial function of the item scores if the items are dichotomous. However, the concept of the configural scale is not restricted to dichotomous items. If the items are polychotomous, the only change is that the number of possible answer patterns will increase.

Theory

A. *The equav model*

The configural scale is defined as the set of answer pattern means, and can be represented by a polynomial function of the item scores. An example of the configural scale and polynomial equation for two items is given in Table 1. In Table 1 the answer patterns are designated by A_0 for the answer

*We are indebted to Professor James G. Taylor for his helpful suggestions.

TABLE I

The Configural Scale and Equav Scores for Two Items

	Answer pattern	Answer pattern frequency	Criterion means	Equav scores			
	Items	n	$\bar{C}$	M			
	1 2			X_0	X_1	X_2	X_{12}
A_0	Y Y	n_0	$\bar{C}_0$	1	1	1	1
A_1	N Y	n_1	$\bar{C}_1$	1	-1	1	-1
A_2	Y N	n_2	$\bar{C}_2$	1	1	-1	-1
A_{12}	N N	n_{12}	$\bar{C}_{12}$	1	-1	-1	1
		Equav regression coefficients		d_0	d_1	d_2	d_{12}

pattern containing all *yes* responses, A_1 for the answer pattern containing a *no* to Item 1 and a *yes* to all other items, A_2 for the answer pattern with a *no* to Item 2 and a *yes* to all other items and in general, A_r where r designates the items to which the subject has responded *no*. The frequency and the criterion mean of each answer pattern is designated by the same subscript; e.g., n_0 and $\bar{C}_0$ are the frequency and criterion mean, respectively, of A_0, the *yes-yes* pattern.

Each item is scored $+1$ for a *yes* response and -1 for a *no* response. Let u_k designate the score for the kth item. Let X_j be a polynomial term, where j indicates the items which form the term. The score on X_j is obtained by actually multiplying the item scores together; e.g., $X_1 = u_1$, $X_{12} = u_1u_2$, $X_{123} = u_1u_2u_3$, and so on. These polynomial terms are called *equav* scores. As will be explained in detail later, *equav* is used to refer to the model for a factorial analysis of variance with equal cell frequencies.

The matrix M as shown in Table 1 is the set of equav scores whose rows are the 2^t answer patterns and whose columns are the 2^t polynomial terms. The general entry, m_{rj}, equals $(-1)^g$, where g is the number of common items in the rth answer pattern and the jth polynomial term. X_0 is a dummy term with a score of $+1$ for every individual, to allow for the constant term in the polynomial equation.

In Table 1 the equav predictor for the criterion score of the ith individual is

$$\hat{C}_i = d_0X_{i0} + d_1X_{i1} + d_2X_{i2} + d_{12}X_{i12}, \tag{1}$$

where $\hat{C}_i$ is the predicted criterion score for the ith individual,
X_{ij} is the score of the ith individual on the jth polynomial term,
d_j is the least squares regression coefficient for the jth polynomial term.

The exact solution for the 2^t regression coefficients can be obtained as follows. Let Z be an N by 2^t matrix whose general element z_{ij} is the equav score of the ith individual on the jth polynomial term. Z is an expanded form of M where the rth row of M is repeated n_r times. Let C be an N-rowed column vector, where c_i is the criterion score of the ith individual. Let d be the $2^t \times 1$ column vector of regression coefficients. Then

$$d = (Z'Z)^{-1}Z'C \tag{2}$$

is the set of regression coefficients which gives the exact least squares fit. The predicted score $\hat{C}_i$ of the ith individual will be the mean of his answer pattern.

Let n be the diagonal matrix of answer pattern frequencies. Then

$$Z'Z = M'nM \tag{3}$$

and

$$Z'C = M'n\bar{C}, \tag{4}$$

where $\bar{C}$ is the 2^t by 1 column vector of answer pattern means. Substituting (3) and (4) in (2),

$$d = (M'nM)^{-1}M'n\bar{C}. \tag{5}$$

Since $M = M'$ and $M^{-1} = 2^{-t}M$,

$$d = M^{-1}n^{-1}M^{-1}Mn\bar{C} = M^{-1}\bar{C} = 2^{-t}M\bar{C}. \tag{6}$$

Thus, each regression coefficient is equal to the algebraic sum of the 2^t criterion averages divided by 2^t.

Scoring each item alternative $+1$ or -1 is exactly analogous to a two-level factorial analysis of variance model when all cells (answer patterns) have equal frequencies. For this reason the term *equav* is used to denote this method of scoring the polynomial terms. In previous papers [1, 3] the items have been scored $+1$ for a yes response and 0 for a no response. This leads to considerable difficulty if item scores are arbitrary. A reversal of an item score involves nonlinear transformations of the polynomial terms which alter the absolute values of the regression coefficients.

Equav scoring has certain algebraic advantages (such as $M = M'$ and $M^{-1} = 2^{-t}M$). The most important advantage is that the absolute values of the regression coefficients are invariant no matter what item scores are reversed. The proof follows:

Reverse the equav score on the kth item. Every -1 becomes $+1$, every $+1$ becomes -1. In other words, $-u_k$ is substituted for u_k. This amounts to multiplying each column of M by -1 if u_k appears in that polynomial term. In general, reversing any set of item scores is equivalent to multiplying the appropriate columns of M by -1.

Let H be a 2^t by 2^t diagonal matrix containing -1 in each diagonal cell corresponding to the appropriate column of M. All other diagonals contain $+1$. Let P be the M matrix after the item scores have been reversed. Then $P = MH$. Let e denote the set of regression coefficients for the polynomial equation using the reversed scores.

Substituting in (5)

$$e = (P'nP)^{-1}P'n\bar{C}; \tag{7}$$

$$e = (H'M'nMH)^{-1}H'M'n\bar{C}; \tag{8}$$

$$e = H^{-1}M^{-1}n^{-1}M^{-1}H^{-1}HMn\bar{C}; \tag{9}$$

$$e = H^{-1}M^{-1}\bar{C}. \tag{10}$$

Since

$$H^{-1} = H \quad \text{and} \quad M^{-1} = 2^{-t}M, \tag{11}$$

$$e = H(2^{-t}M\bar{C}). \tag{12}$$

Therefore

$$e = Hd. \tag{13}$$

Premultiplying d by H simply reverses the sign of certain of the d coefficients. This proves that the absolute values of the equav coefficients are invariant under reversal of item scores.

So far, only matters of algebraic and computational convenience have been discussed. However, it is possible that certain methods of scoring items may be most appropriate with certain kinds of test content; i.e., equav scoring may be most appropriate for personality tests, and zero-one scoring may be most appropriate to aptitude and achievement tests.

Suppose that certain of the 2^t regression coefficients are zero in the population. Then (6) does not give exact least square estimates of the nonzero coefficients for the sample. A general solution for this case where certain coefficients are assumed to be zero has been given in ([3], equation 38).

As an example, consider the special case of linearity where all the coefficients for the nonlinear terms are zero. Then

$$\hat{C} = w_0 + w_1X_1 + w_2X_2 + w_3X_3 + w_4X_4\ . \tag{14}$$

Let Z_t be the N by t submatrix formed by taking the first $t + 1$ columns of Z. Then

$$w_t = (Z_t'Z_t)^{-1}Z_t'C, \tag{15}$$

where w_t is the column of $t + 1$ linear regression coefficients. Let K_t be the 2^t by $t + 1$ submatrix formed by taking the first $t + 1$ columns of M. Then $Z_t'Z_t = K_t'nK_t$ and $Z_t'C = K_t'n\bar{C}$. Therefore

$$w_t = (K_t' n K_t)^{-1} K_t' n \bar{C}. \tag{16}$$

Note that since K_t is rectangular, no simple inverse exists and equation (16) cannot be further simplified.

B. Restrictions on the answer pattern means

Any method of scoring the t items which yields optimal validity in the population and yet uses fewer than 2^t parameters imposes certain restrictions on the population answer pattern means. In this paper, restrictions imposed by three well known test scoring methods: multiple regression, total score, and multiple cut-off, are considered.

Table 2 summarizes the necessary and sufficient conditions (in the mathematical sense) for each of the three scoring techniques to yield optimal validity. The restrictions on the equav coefficients amount to definitions in the case of multiple regression and total score. From these definitions a number of restrictions on the answer pattern means can be derived.

TABLE 2

Conditions for Optimal Validity

	Scoring Method		
	Multiple Regression	Total Score	Multiple Cut-off
Equav Coefficients	All non-linear coefficients are zero	1. All non-linear coefficients are zero 2. All first-order coefficients are equal	Only one coefficient differs from the others in absolute value
Answer Pattern Means	The sum of complementary answer pattern means is equal to a constant, $2d_0$	1. The sum of complementary answer pattern means is equal to a constant, $2d_0$ 2. All answer patterns whose sums of item scores are equal have equal means.	Only one mean differs from the others.

One of the most useful restrictions on the answer pattern means in the linear case is given in Table 2. First, let us define *complementary* answer patterns. Answer pattern A_r is complementary to $A_{r'}$, if and only if, every item response in A_r is reversed for $A_{r'}$. For example A_{12} (NNYY) is the complement of A_{34} (YYNN). In our four-item case

$$\bar{C}_r = d_0 + d_1X_1 + d_2X_2 + d_3X_3 + d_4X_4 , \tag{17}$$

$$\bar{C}_{r'} = d_0 + d_1X_1' + d_2X_2' + d_3X_3' + d_4X_4' . \tag{18}$$

For the equav model, $X_1 + X_1' = X_2 + X_2' = X_3 + X_3' = X_4 + X_4' = 0$. In general $(X + X') = 0$. Therefore,

$$\bar{C}_{r'} + \bar{C}_r = 2d_0 + (X + X')\,(d_1 + d_2 + d_3 + d_4) = 2d_0\,. \tag{19}$$

The additional restriction for the total score case, that all answer patterns with the same total score have equal means, can be derived as follows: By definition, the first-order coefficients are equal, i.e.,

$$d_1 = d_2 = d_3 = d_4 = d.$$

Therefore,

$$\begin{aligned} \bar{C}_r &= d_0 + d_1X_1 + d_2X_2 + d_3X_3 + d_4X_4 \\ &= d_0 + d(X_1 + X_2 + X_3 + X_4). \end{aligned} \tag{20}$$

Thus all answer patterns with the same sum of item scores $(X_1 + X_2 + X_3 + X_4)$ will have the same mean.

The basic definition of the multiple cut-off is that only two scores are used. The subjects in the all-yes answer pattern are assigned one score; the subjects in all other answer patterns are assigned the other score. This scoring method implies that for optimal validity all except one of the answer pattern means should be equal. Without losing generality, it can be assumed that the unique mean is $\bar{C}_0$. Let $\bar{C}$ denote the constant means for all other answer patterns. From (6) for calculating the regression coefficients from the means it follows that

$$d_0 = [\bar{C}_0 + (2^t - 1)\bar{C}]/2^t, \tag{21}$$

$$d_1 = (\bar{C}_0 - \bar{C})/2^t, \tag{22}$$

$$d_2 = (\bar{C}_0 - \bar{C})/2^t, \tag{23}$$

and in general

$$d_j = (\bar{C}_0 - \bar{C})/2^t. \tag{24}$$

Thus, in the multiple cut-off case, all coefficients but one will have the same absolute value.

C. The F ratio tests

How any hypothesized relation of a specific set of item interactions to the criterion can be tested by means of the F ratio is shown in [3]. Of course, the usual assumptions of normality and homogeneity of variance must be met.

The general F ratio test is as follows: let η_c be the configural validity, r_0 be the validity of any specified scoring method, and v_0 be the number of sample statistics that must be calculated. Then

$$F = \left(\frac{\eta_c^2 - r_0^2}{1 - \eta_c^2}\right)\left(\frac{N - 2^t}{2^t - v_0}\right). \tag{25}$$

An even more general formula is given in ([3], equation 34).

Worked Example

In order to illustrate the method an example was constructed. Five hundred scores were drawn from a table of normal random deviates. These scores were then transformed so that the universe mean was 5 and the universe standard deviation was 10. The frequencies for each answer pattern were calculated by fixing the p values of the four items at $p_1 = .3$, $p_2 = .4$ $p_3 = .5$, $p_4 = .6$ and assuming all items to be statistically independent. The 500 scores were assigned at random to the sixteen answer patterns according to predetermined frequencies.

To the artificial data described above, a linear systematic component was added. The following arbitrary values were assigned: $d_0 = 22$, $d_1 = -1$, $d_2 = 5$, $d_3 = -7$, $d_4 = 9$. Using these values in (14) gave the systematic component that was added to each answer pattern mean. The column in Table 3 labelled $\bar{C}$ gives the answer pattern mean obtained by these computations.

TABLE 3

Basic Data for Worked Example

Answer Pattern		n	ΣC	ΣC^2	$\frac{1}{N}(\Sigma C)^2$	Deviance	$\bar{C}$	d	$\hat{C}$	T	M
A_0	YYYY	18	605	22,581	20,334.722	2,246.277	33.611	26.423	33.465	2	0
A_1	NYYY	42	1,556	61,114	57,646.095	3,467.905	37.048	-1.523	36.397	3	0
A_2	YNYY	27	604	17,448	13,511.703	3,936.296	22.370	5.145	23.011	1	0
A_3	YYNY	18	861	42,275	41,184.500	1,090.500	47.833	-6.230	45.745	3	1
A_4	YYYN	12	168	3,836	2,352.000	1,484.000	14.000	9.541	14.679	1	0
A_{12}	NNYY	63	1,725	52,863	47,232.142	5,630.857	27.381	-.121	25.943	2	0
A_{13}	NYNY	42	1,997	100,023	94,952.595	5,070.405	47.548	-.007	48.677	4	0
A_{14}	NYYN	28	505	11,821	9,108.035	2,712.964	18.036	.282	17.611	2	0
A_{23}	YNNY	27	947	35,765	33,215.148	2,549.852	35.074	.336	35.291	2	0
A_{24}	YNYN	18	84	1,972	392.000	1,580.000	4.667	.402	4.255	0	0
A_{34}	YYNN	12	291	7,813	7,056.750	756.250	24.250	.369	26.959	2	0
A_{123}	NNNY	63	2,321	91,363	85,508.587	5,854.413	36.841	-.217	38.223	3	0
A_{124}	NNYN	42	186	5,738	823.714	4,914.286	4.429	.574	7.157	1	0
A_{134}	NYNN	28	846	27,548	25,561.285	1,986.714	30.214	-.863	29.891	3	0
A_{234}	YNNN	18	313	6,495	5,442.722	1,052.278	17.389	-.656	16.505	1	0
A_{1234}	NNNN	42	927	24,471	20,460.214	4,010.786	22.071	.157	19.437	2	0
	Total	500	13,936	513,126	388,424.192	124,701.808	27.872	33.612	423.216		

$\Sigma\hat{C} = 13,935.300$ $\Sigma\hat{C}^2 = 463,601.606$ $\Sigma\hat{C}C = 463,603.728$

$\Sigma M = 18$ $\Sigma M^2 = 18$ $\Sigma MC = 861$ $\Sigma T = 1100$ $\Sigma T^2 = 2,890$ $\Sigma TC = 36,011$

Step 1—*Calculation of the configural validity*

First, a one-way analysis of variance is computed. The column labelled *deviance* in Table 3 contains the sum of squared deviations (sum of squares) about each answer pattern mean. The sum of the 16 answer pattern deviances is W, the within group sum of squares. The column labelled $(\sum C)^2/N$ contains a correction term for each answer pattern. The sum of the 16 correction terms minus the correction for the total equals B, the between group sum of squares.

TABLE 4

Analysis of Variance

Source	df	Deviance	Mean Square
Between Answer Patterns	15	76,358.020	5,090.535
Within	484	48,343.784	99.884
Total	499	124,701.804	
$\eta_c^2 = .612$		F = 50.964	

These figures along with T, the deviance (sum of squares) about the total mean, are given, in the usual analysis of variance form, in Table 4. The formula for the configural validity is

$$\eta_c^2 = B/T = \frac{76,\,358.020}{124,\,701.804} = .612325. \tag{26}$$

The test of significance is

$$F = \left(\frac{\eta_c^2}{1 - \eta_c^2}\right)\left(\frac{N - 2^t}{2^t - 1}\right) = \left(\frac{.612325}{.387675}\right)\left(\frac{484}{15}\right) = 50.964. \tag{27}$$

Since the .001 confidence level is 2.577, the configural validity is obviously greater than zero. If the F ratio was insignificant, the analysis would be stopped, for then no method of scoring the test would give a better-than-chance prediction of the criterion scores.

Step 2—*Calculation of the polynomial regression coefficients*

In Table 3, the column labelled d contains all 16 polynomial regression coefficients. Each coefficient was computed by adding together the 16 means (appropriately signed) and dividing by 2^t. The sign of each mean is given by $(-1)^g$, where g is the number of common items in the subscripts of the regression coefficient and the mean.

For example,

$$d_0 = \frac{1}{2^t} [\bar{C}_0 + \bar{C}_1 + \bar{C}_2 + \bar{C}_3 + \bar{C}_4 + \bar{C}_{12} + \bar{C}_{13} + \bar{C}_{14} + \bar{C}_{23} + \bar{C}_{24}$$

$$+ \bar{C}_{34} + \bar{C}_{123} + \bar{C}_{124} + \bar{C}_{134} + \bar{C}_{234} + \bar{C}_{1234}] = \frac{422.762}{16} = 26.423,$$

$$d_1 = \frac{1}{2^t} [\bar{C}_0 - \bar{C}_1 + \bar{C}_2 + \bar{C}_3 + \bar{C}_4 - \bar{C}_{12} - \bar{C}_{13} - \bar{C}_{14} + \bar{C}_{23} + \bar{C}_{24}$$

$$+ \bar{C}_{34} - \bar{C}_{123} - \bar{C}_{124} - \bar{C}_{134} + \bar{C}_{234} - \bar{C}_{1234}] = \frac{-24.374}{16} = -1.523,$$

$$d_{12} = \frac{1}{2^t} [\bar{C}_0 - \bar{C}_1 - \bar{C}_2 + \bar{C}_3 + \bar{C}_4 + \bar{C}_{12} - \bar{C}_{13} - \bar{C}_{14} - \bar{C}_{23} - \bar{C}_{24}$$

$$+ \bar{C}_{34} + \bar{C}_{123} + \bar{C}_{124} - \bar{C}_{134} - \bar{C}_{234} + \bar{C}_{1234}] = \frac{-1.930}{16} = -.121,$$

and so on.

These coefficients can be scanned in order to see which scoring method seems to give an optimal prediction of the criterion with the fewest parameters. For example, if, in the population, the relation between the items and the criterion were exactly linear, the only nonzero coefficients would be d_0, d_1, d_2, d_3, and d_4. In any actual sample, the other nonlinear coefficients would be small but not exactly zero. So one can simply look at the five linear coefficients to see if their absolute values are larger than any of the other coefficients. Another such test is to check the frequency of negative values among the eleven non linear coefficients. In the linear case, the true probability of a negative value is 1/2. If these crude combinatorial tests do not contradict the hypothesis of linearity we proceed to the next possibility—that the total score will give maximum validity.

The total score will give maximum validity in the population when all the conditions for linearity are met and in addition, $| d_1 | = | d_2 | = | d_3 | = | d_4 |$; i.e., the absolute values of the first-order coefficients are equal. Again, a crude check of this can be made in the sample by seeing if there is wide variation among the absolute values of the first-order coefficients.

In Table 3 it can be seen that the linear coefficients (d_0, d_1, d_2, d_3, and d_4) meet the first condition; their absolute values are larger than those of the nonlinear coefficients. Also the second condition is met; the ratio of negative nonlinear coefficients was 4/11, which is not significantly different from the expected value of 1/2. Therefore, the hypothesis of linearity is not contradicted.

Next the first-order coefficients were examined to see if the total score was likely to have maximum validity. If so, the first-order coefficients (d_1,

d_2 , d_3 , and d_4) would be approximately equal. But d_4 was more than six times the absolute value of d_1 . So it is unlikely that total score would have maximum validity.

As mentioned earlier the above crude tests can be used if the research worker has no definite hypothesis about the optimal scoring method. However, to demonstrate conclusively that linear scoring is sufficient for optimal brediction it is necessary to show that the multiple correlation, $R_{c.1,2,3,4}$, does not differ significantly from the configural validity η_c . To demonstrate conclusively that linear scoring is preferable to the total score and the multiple cut-off score, it has to be shown that the total score validity and the multiple cut-off validity are significantly less than the configural validity.

Step 3—Calculation of the multiple correlation

In Step 2, the linear hypothesis passed the first crude tests. To compute the linear multiple correlation $R_{c.1,2,3,4}$ first $\hat{C}_r$, the predicted criterion score for the rth answer pattern, was calculated. (This may not be the most convenient method, but it does show any large deviations from the linear hypothesis. For a perfect linear fit, $\hat{C}_r = \bar{C}_r$.) To obtain $\hat{C}$ it was first necessary to compute the linear regression coefficients; i.e., w_0 , w_1 , w_2 , w_3 , w_4 from (16).

The matrices $(K_t'nK_t)$, $(K_t'nK_t)^{-1}$ and $(K_t'n\bar{C})$ are presented in Table 5. The regression coefficients are in the column w. Then $\hat{C}$ was obtained by applying the equation $\hat{C} = K_t w$. The predicted criterion means are presented in the $\hat{C}$ column of Table 3. $R_{c.1,2,3,4}$ is equal to $r_{c\hat{c}}$, the zero-order correlation between C and $\hat{C}$.

TABLE 5

Computation of Linear Regression Coefficients

	$K_t'nK_t$						$1000(K_t'nK_t)^{-1}$					$K_t'n\bar{C}$	w
	x_0	x_1	x_2	x_3	x_4		x_0	x_1	x_2	x_3	x_4		
x_0	500	-200	-100	0	100	x_0	2.548	.952	.417	0	-.417	13,936	26.451
x_1	-200	500	40	0	-40	x_1	.952	2.381	0	0	0	-6,190	-1.466
x_2	-100	40	500	0	-20	x_2	.417	0	2.083	0	0	-278	5.227
x_3	0	0	0	500	0	x_3	0	0	0	2.000	0	-3,070	-6.140
x_4	100	-40	-20	0	500	x_4	-.417	0	0	0	2.083	7,296	9.393
Sum	300	300	420	500	540		3.500	3.333	2.500	2.000	1.666	11,694	33.465

This was computed by the well known formula

$$r_{c\hat{c}}^2 = \frac{(N \sum C\hat{C} - \sum C \sum \hat{C})^2}{[N \sum C^2 - (\sum C)^2][N \sum \hat{C}^2 - (\sum \hat{C})^2]}. \tag{28}$$

Column $\sum C$ in Table 3 contains the sums of the criterion scores for each answer pattern. To obtain $\sum C\hat{C}$, each $\hat{C}_r$ was multiplied by the $\sum C$ for the rth answer pattern, and the result was summed over all patterns. Similarly,

$$\sum \hat{C}^2 = \sum_{r=1}^{2^t} n_r \hat{C}_r^2 = \sum \hat{C}C \quad \text{and} \quad \sum \hat{C} = \sum_{r=1}^{2^t} n_r \hat{C}_r = \sum C.$$

Substituting in (28) from the data in Table 3,

$$r_{c\hat{c}}^2 = \frac{[500(463,603.728) - 13,936(13,935.300)]^2}{[500(513,126) - (13,936)^2][500(463,601.606) - (13,935.300)^2]} = .603.$$

Step 4—Comparison of the multiple correlation with the configural validity

Applying (25),

$$F = \left(\frac{.612 - .603}{.388}\right)\left(\frac{500 - 16}{16 - 5}\right) = 1.021.$$

Since, for 11 and 484 d.f., the .05 level is 1.750, the F test indicates that the multiple correlation does not differ significantly from the configural validity; i.e., the multiple regression scoring method yields optimal validity.

Step 5—Calculation of the total score validity

In order to rule out conclusively the total score hypothesis, the zero-order correlation r_t was computed between the total score and the criterion. In general, the total score, T, is equal to the number of yes responses for all items with positive first-order coefficients plus the number of no responses for all items with negative first-order coefficients. The column labelled T in Table 3 gives the total score for each answer pattern. The usual formula for the squared correlation was used.

$$(29) \qquad r_t^2 = \frac{(N \sum CT - \sum C \sum T)^2}{[(N \sum C^2 - (\sum C)^2][N \sum T^2 - (\sum T)^2]}.$$

Using the data from Table 3,

$$r_t^2 = \frac{[500(36,011) - (13,936)(1,100)]^2}{[500(513,126) - (13,936)^2][500(2,890) - (1,100)^2]} = .489.$$

Step 6—Comparison of the total score validity with the configural validity

Applying (25),

$$F = \left(\frac{.612 - .489}{.388}\right)\left(\frac{500 - 16}{16 - 2}\right) = 10.960.$$

Since, for 14 and 484 d.f., the .05 level is 1.690, the F test indicates that the total score validity is significantly less than the configural validity, i.e., the total score does not yield optimal validity.

Step 7—Calculation of the multiple cut-off validity

In order to rule out conclusively the multiple cut-off hypothesis, the zero-order correlation, r_{mc} , was computed between the multiple cut-off score and the criterion. Multiple cut-off scoring demands that a score of one be assigned to the answer pattern with the highest (or lowest) mean and that a score of zero be assigned to all other answer patterns. Column M in Table 3 gives the multiple cut-off score for each answer pattern. Substituting figures from Table 3 into the formula for squared correlation,

$$r_{mc}^2 = \frac{[500(861) - 18(13{,}936)]^2}{[500(513{,}126) - (13{,}936)^2][500(18) - (18)^2]} = .060.$$

Step 8—Comparison of the multiple cut-off validity with the configural validity

Applying (25),

$$F = \left(\frac{.612 - .060}{.388}\right)\left(\frac{500 - 16}{16 - 2}\right) = 49.185.$$

Since, for 14 and 484 d.f., the .05 confidence level is 1.690, the F test shows that the multiple cut-off validity is significantly less than the configural validity.

Discussion

It has been shown how the concept of the configural scale can be used to give an exact statistical test of whether a selected scoring technique has optimal validity. Worked examples have been given for three well known test scoring methods: multiple regression, multiple cut-off, and total score. In general, the principal advantage of configural analysis is that *all* of the information concerning the subject's test behavior is utilized.

On the other hand, the principal disadvantages of configural analysis lie in the very fact that all the information is conserved; i.e., all possible answer patterns are considered. In a t-item test the formula for the configural validity involves 2^t parameters; i.e., 2^t answer pattern averages. It is immediately obvious that this technique is only appropriate for situations where the number of items is very small compared to the number of subjects—N must be much greater than 2^t. For example, even when the number of items is as small as 10, 2^t will be 1024.

Use of the equav coefficients for scanning purposes introduces another difficulty. The F ratio test no longer gives the exact confidence level; it is simply a decision function. The procedure of selecting the test scoring method which is most likely to yield optimal validity alters the significance level of the F test (cf. [2], p. 199 ff.). As a way of deciding among several possible test scoring methods, the scanning technique is certainly a reasonable procedure. However, it is advisable, after selecting a test scoring method on one sample, to cross-validate it on another sample.

Configural analysis is most suitable in situations where testing time is short and the number of subjects is large. For example, take the case of neuropsychiatric screening in the armed forces where often only a few minutes of testing time is available, and a very large number of subjects must be screened. Here, items should be constructed in such a way that all 2^t regression coefficients are significant. This will give maximum discrimination. However, in actual practice some of the regression coefficients will probably be nonsignificant. If this occurs, a value of zero should be given to all nonsignificant coefficients. The use of any other values will lower the validity of the test.

REFERENCES

[1] Horst, P. Pattern analysis and configural scoring. *J. clin. Psychol.*, 1954, **10**, 3-11.

[2] Kendall, M. G. The advanced theory of statistics. Vol. II. London: Griffin, 1948.

[3] Lubin, A. and Osburn, H. G. A theory of pattern analysis for the prediction of a quantitative criterion. *Psychometrika*, 1957, 22, 63-73.

Manuscript received 12/7/56

Revised manuscript received 4/18/57

PSYCHOMETRIKA—VOL. 22, NO. 4
DECEMBER, 1957

A MODEL FOR RESPONSE TENDENCY COMBINATION*

DAVID BIRCH

UNIVERSITY OF MICHIGAN

A model is proposed to predict the performance on a compound stimulus as a function of the performance on the component stimuli in a two-choice situation. Data from a learning task are used to evaluate the model.

Any theory of behavior which analyzes a stimulus complex into components and attempts to account for responses to the complex on the basis of response tendencies to the components faces the problem of specifying the rule for the combination of the component response tendencies. Theorists such as Hull [4], Thurstone [8], Gulliksen [3], Spence [7], Estes and Burke [2], Bush and Mosteller [1], and Restle [5] have incorporated combination rules within their theories and then made use of them in deriving implications from their theories. Seldom, however, has the combination rule itself been the focus of attention for these theorists. One recent instance is a study by Schoeffler [6] who carried out a test of a combination rule derived from the Estes-Burke learning theory. The rule is linked directly to the parameters of the theory and certain assumptions about the parameters are made by Schoeffler in bringing the rule to test.

This paper presents the development of a model for combining response tendencies in a two-choice situation and reports a test for the fit of the model to data. The basis for the definition of the parameters of the model proposed, as well as the impetus for the development of the model, are derived from Hullian behavior theory. However, the combination rule, specified by the interrelationships of the parameters of the model, does not depend upon any particular learning theory and, therefore, may be of value in a variety of situations where problems of combination arise.

A Model for Response Tendency Combination

The difference in response tendency strength for stimulus a, $D_a = ({}_aE_u - {}_aE_s)$, at any given point in time will be considered to be in one of three states: $D_a \geq d$, $D_a \leq -d$, or $-d < D_a < d$, where d is a parameter with a value such that $\Pr\{u|D_a \geq d\} = 1$, $\Pr\{u|D_a \leq -d\} = 0$, and $\Pr\{u|\ -d < D_a < d\} = .5$. Let $\Pr\{D_a \geq d\}$, $\Pr\{D_a \leq -d\}$, and $\Pr\{-d <$

*The initial work on the model was carried out under the Summer Faculty Research Fellowship Program of the Horace H. Rackham School of Graduate Studies at the University of Michigan. Acknowledgment is also due Mr. Richard Anderson for his assistance in data collection.

$D_a < d\} = 1 - \Pr \{D_a \geq d\} - \Pr \{D_a \leq -d\}$ be the probabilities that the difference in response tendency strengths is in each of the three states.

It then follows that the compound probability of obtaining u and v when a is presented may be written as

$$\Pr \{u|a\} = \Pr \{D_a \geq d\} + (.5)[1 - \Pr \{D_a \geq d\} - \Pr \{D_a \leq -d\}] \tag{1}$$

and

$$\Pr \{v|a\} = \Pr \{D_a \leq -d\} + (.5)[1 - \Pr \{D_a \geq d\} - \Pr \{D_a \leq -d\}]. \tag{2}$$

A corresponding development for b gives

$$\Pr \{u|b\} = \Pr \{D_b \geq d\} + (.5)[1 - \Pr \{D_b \geq d\} - \Pr \{D_b \leq -d\}] \tag{3}$$

and

$$\Pr \{v|b\} = \Pr \{D_b \leq -d\} + (.5)[1 - \Pr \{D_b \geq d\} - \Pr \{D_b \leq -d\}]. \tag{4}$$

Since $\Pr \{u|a\} + \Pr \{v|a\} = 1$ and $\Pr \{u|b\} + \Pr \{v|b\} = 1$, there are available two independent equations in the four unknowns, $\Pr \{D_a \geq d\}$, $\Pr \{D_a \leq -d\}$, $\Pr \{D_b \geq d\}$, and $\Pr \{D_b \leq -d\}$.

The compound stimulus (a, b) is defined as the joint presentation of a and b, and this compound stimulus can be characterized in four mutually exclusive and exhaustive ways by the responses obtained to a and b upon separate presentation of these stimuli. That is, a and b may both be responded to with u, a may be responded to with u and b with v, a may be responded to with v and b with u, or both a and b may be responded to with v. Let the corresponding designations of (a, b) be (a_u, b_u), (a_u, b_v), (a_v, b_u), and (a_v, b_v).

If the total probability of u to the presentation of (a, b) is denoted $\Pr \{u|(a, b)\}$, then

$$\begin{aligned} \Pr \{u|(a, b)\} = {} & \Pr \{u|(a_u, b_u)\} \cdot \Pr \{a_u, b_u\} + \Pr \{u|(a_u, b_v)\} \\ & \cdot \Pr \{a_u, b_v\} + \Pr \{u|(a_v, b_u)\} \cdot \Pr \{a_v, b_u\} \\ & + \Pr \{u|(a_v, b_v)\} \cdot \Pr \{a_v, b_v\}, \end{aligned} \tag{5}$$

where the entries on the right-hand side of the equation are the independent contributions from the four classes of (a, b). By writing each of these terms separately as a function of $\Pr \{D_a \geq d\}$, $\Pr \{D_a \leq -d\}$, $\Pr \{D_b \geq d\}$, and $\Pr \{D_b \leq -d\}$, a total of six experimentally independent equations in the four unknowns will be available so that the values of the unknowns are overdetermined.

It follows from (1), (2), (3), and (4) that the probabilities of occurrence of the four classes of (a, b) are

$$\begin{aligned}\Pr\{a_u, b_u\} &= \Pr\{u|a\}\cdot\Pr\{u|b\} = \Pr\{D_a \geq d\}\cdot\Pr\{D_b \geq d\}\\ &\quad + \Pr\{D_a \geq d\}\cdot(.5)[1 - \Pr\{D_b \geq d\} - \Pr\{D_b \leq -d\}]\\ &\quad + (.5)[1 - \Pr\{D_a \geq d\} - \Pr\{D_a \leq -d\}]\cdot\Pr\{D_b \geq d\}\\ &\quad + (.5)[1 - \Pr\{D_a \geq d\} - \Pr\{D_a \leq -d\}]\\ &\quad \cdot(.5)[1 - \Pr\{D_b \geq d\} - \Pr\{D_b \leq -d\}];\end{aligned} \tag{6}$$

$$\begin{aligned}\Pr\{a_u, b_v\} &= \Pr\{u|a\}\cdot\Pr\{v|b\} = \Pr\{D_a \geq d\}\\ &\quad \cdot\Pr\{D_b \leq -d\} + \Pr\{D_a \geq d\}\cdot(.5)[1 - \Pr\{D_b \geq d\}\\ &\quad - \Pr\{D_b \leq -d\}] + (.5)[1 - \Pr\{D_a \geq d\}\\ &\quad - \Pr\{D_a \leq -d\}]\cdot\Pr\{D_b \leq -d\}\\ &\quad + (.5)[1 - \Pr\{D_a \geq d\} - \Pr\{D_a \leq -d\}]\\ &\quad \cdot(.5)[1 - \Pr\{D_b \geq d\} - \Pr\{D_b \leq -d\}];\end{aligned} \tag{7}$$

$$\begin{aligned}\Pr\{a_v, b_u\} &= \Pr\{v|a\}\cdot\Pr\{u|b\} = \Pr\{D_a \leq -d\}\cdot\Pr\{D_b \geq d\}\\ &\quad + \Pr\{D_a \leq -d\}\cdot(.5)[1 - \Pr\{D_b \geq d\} - \Pr\{D_b \leq -d\}]\\ &\quad + (.5)[1 - \Pr\{D_a \geq d\} - \Pr\{D_a \leq -d\}]\cdot\Pr\{D_b \geq d\}\\ &\quad + (.5)[1 - \Pr\{D_a \geq d\} - \Pr\{D_a \leq -d\}]\\ &\quad \cdot(.5)[1 - \Pr\{D_b \geq d\} - \Pr\{D_b \leq -d\}];\end{aligned} \tag{8}$$

and

$$\begin{aligned}\Pr\{a_v, b_v\} &= \Pr\{v|a\}\cdot\Pr\{v|b\} = \Pr\{D_a \leq -d\}\cdot\Pr\{D_b \leq -d\}\\ &\quad + \Pr\{D_a \leq -d\}\cdot(.5)[1 - \Pr\{D_b \geq d\} - \Pr\{D_b \leq -d\}]\\ &\quad + (.5)[1 - \Pr\{D_a \geq d\} - \Pr\{D_a \leq -d\}]\cdot\Pr\{D_b \leq -d\}\\ &\quad + (.5)[1 - \Pr\{D_a \geq d\} - \Pr\{D_a \leq -d\}]\\ &\quad \cdot(.5)[1 - \Pr\{D_b \geq d\} - \Pr\{D_b \leq -d\}].\end{aligned} \tag{9}$$

The probability of u for each of the classes may be obtained by weighting each component of $\Pr\{a_u, b_u\}$, $\Pr\{a_u, b_v\}$, $\Pr\{a_v, b_u\}$, and $\Pr\{a_v, b_v\}$ by an appropriate conditional probability of occurrence of u. The weights assumed are as follows: the conditional probability of u is 1, given that the combinations of response tendency states for a and b are $D_a \geq d$ and $D_b \geq d$, or $D_a \geq d$ and $-d < D_b < d$, or $-d < D_a < d$ and $D_b \geq d$; the conditional probability of u is 0 given $D_a \leq -d$ and $D_b \leq -d$, or $D_a \leq -d$ and $-d < D_b < d$, or $-d < D_a < d$ and $D_b \leq -d$; and the conditional probability of u is .5 given $D_a \geq d$ and $D_b \leq -d$, or $D_a \leq -d$ and $D_b \geq d$, or $-d <$

TABLE 1

Assumed Conditional Probability of Occurrence of u to the Compound Stimulus for the Possible Combinations of Response Tendency States of a and b.

Response Tendency States for a	Response Tendency States for b		
	$(D_b \geq d)$	$(-d < D_b < d)$	$(D_b \leq -d)$
$(D_a \geq d)$	1	1	.5
$(-d < D_a < d)$	1	.5	0
$(D_a \leq -d)$	.5	0	0

$D_a < d$ and $-d < D_b < d$. These assumed values are presented in Table 1.

These weights in conjunction with (6), (7), (8), and (9) produce four equations in the four unknowns Pr $\{D_a \geq d\}$, Pr $\{D_a \leq -d\}$, Pr $\{D_b \geq d\}$, and Pr $\{D_b \leq -d\}$. Since the model under development was instigated by the problem of the prediction of performance to a compound stimulus as a function of the performance to the component stimuli, the relationships of (1), (2), (3), and (4) may be used to reduce (6), (7), (8), and (9) to functions of the two unknowns Pr $\{D_a \geq d\}$ and Pr $\{D_b \geq d\}$. The resulting, simplified equations are

$$\begin{aligned}\Pr\{u|(a_u, b_u)\}\cdot\Pr\{a_u, b_u\} &= (.5)[\Pr\{D_a \geq d\}\cdot\Pr\{u|b\} \\ &\quad + \Pr\{D_b \geq d\}\cdot\Pr\{u|a\} - \Pr\{D_a \geq d\}\cdot\Pr\{D_b \geq d\} \\ &\quad + \Pr\{u|a\}\cdot\Pr\{u|b\}];\end{aligned} \tag{10}$$

$$\begin{aligned}\Pr\{u|(a_u, b_v)\}\cdot\Pr\{a_u, b_v\} &= (.5)[\Pr\{D_a \geq d\}\cdot\Pr\{v|b\} \\ &\quad - \Pr\{D_b \geq d\}\cdot\Pr\{u|a\} + \Pr\{u|a\}\cdot\Pr\{u|b\}];\end{aligned} \tag{11}$$

$$\begin{aligned}\Pr\{u|(a_v, b_u)\}\cdot\Pr\{a_v, b_u\} &= (.5)[\Pr\{D_b \geq d\}\cdot\Pr\{v|a\} \\ &\quad - \Pr\{D_a \geq d\}\cdot\Pr\{u|b\} + \Pr\{u|a\}\cdot\Pr\{u|b\}];\end{aligned} \tag{12}$$

and

$$\begin{aligned}\Pr\{u|(a_v, b_v)\}\cdot\Pr\{a_v, b_v\} &= (.5)[-\Pr\{D_a \geq d\} \\ &\quad \cdot\Pr\{u|b\} - \Pr\{D_b \geq d\}\cdot\Pr\{u|a\} + \Pr\{D_a \geq d\} \\ &\quad \cdot\Pr\{D_b \geq d\} + \Pr\{u|a\}\cdot\Pr\{u|b\}].\end{aligned} \tag{13}$$

Finally, (5) becomes:

$$\begin{aligned}\Pr\{u|(a, b)\} &= \Pr\{D_a \geq d\}[(.5) - \Pr\{u|b\}] \\ &\quad + \Pr\{D_b \geq d\}[(.5) - \Pr\{u|a\}] + 2\Pr\{u|a\}\cdot\Pr\{u|b\}.\end{aligned} \tag{14}$$

It may also be noted from (10) and (13) that

$$\begin{aligned}2\,\mathrm{Pr}\,\{u|(a_u\,,\,b_u)\}\cdot\mathrm{Pr}\,\{a_u\,,\,b_u\} - \mathrm{Pr}\,\{u|a\}\cdot\mathrm{Pr}\,\{u|b\} \\ = \mathrm{Pr}\,\{D_a \geq d\}\cdot\mathrm{Pr}\,\{u|b\} + \mathrm{Pr}\,\{D_b \geq d\}\cdot\mathrm{Pr}\,\{u|a\} \\ - \mathrm{Pr}\,\{D_a \geq d\}\cdot\mathrm{Pr}\,\{D_b \geq d\} = \mathrm{Pr}\,\{u|a\}\cdot\mathrm{Pr}\,\{u|b\} \\ - 2\,\mathrm{Pr}\,\{u|(a_v\,,\,b_v)\}\cdot\mathrm{Pr}\,\{a_v\,,\,b_v\},\end{aligned}$$

which indicates that it is necessary that

$$\begin{aligned}\mathrm{Pr}\,\{u|a\}\cdot\mathrm{Pr}\,\{u|b\} - \mathrm{Pr}\,\{u|(a_u\,,\,b_u)\} \\ \cdot\mathrm{Pr}\,\{a_u\,,\,b_u\} = \mathrm{Pr}\,\{u|(a_v\,,\,b_v)\}\cdot\mathrm{Pr}\,\{a_v\,,\,b_v\}\end{aligned}$$

if these two equations are to be consistent. The latter relationship provides a partial test of the model since all four of the values are experimentally independent observables. If this relationship can be shown to hold within reasonable limits, then (10) and (13) may be combined into

$$\begin{aligned}\mathrm{Pr}\,\{u|(a_u\,,\,b_u)\}\cdot\mathrm{Pr}\,\{a_u\,,\,b_u\} - \mathrm{Pr}\,\{u|(a_v\,,\,b_v)\} \\ \cdot\mathrm{Pr}\,\{a_v\,,\,b_v\} = \mathrm{Pr}\,\{D_a \geq d\}\cdot\mathrm{Pr}\,\{u|b\} + \mathrm{Pr}\,\{D_b \geq d\} \\ \cdot\mathrm{Pr}\,\{u|a\} - \mathrm{Pr}\,\{D_a \geq d\}\cdot\mathrm{Pr}\,\{D_b \geq d\}.\end{aligned} \tag{15}$$

A Test of the Model

To obtain data for a test of the model, a learning task was carried out in which subjects were required to associate the response *Dac* to each of ten letter pairs and ten number pairs, and the response *Jix* to each of another set of ten letter pairs and ten number pairs. In dealing with the resulting data, it is convenient to define u as a correct response, C, and v as an incorrect response, I. Also, stimulus a is defined as the set of twenty letter pairs, L, and stimulus b as the set of twenty number pairs, N.

In constructing the letter pairs a total of ten letters (B, F, G, H, K, N, Q, S, Y, and Z) were used, and these were paired in such a fashion that each letter appeared in the first position twice, once to be associated with *Dac* and once with *Jix*, and in the second position twice, again once to be associated with *Dac* and once with *Jix*. No two letters were ever paired more than once. The ten digits (0 through 9) were treated in similar manner.

The subjects, 145 male volunteers from the introductory psychology class, participated in groups of approximately 14. To provide an opportunity for learning, the letter pairs and number pairs were shown individually in a random sequence on flash cards for 10 seconds, with the correct response exposed during the last 6 seconds. Subjects recorded no responses during the learning series but did record their responses during the test series, which was alternated with the learning series.

Three learning and three test series were given. In the test series each of the 20 letter pairs, the 20 number pairs, and 20 compounds were presented individually in random order. Each compound was made up of a letter pair and a number pair chosen at random without replacement under the restriction that the same response be correct for both the letter pair and the number pair. Each test series employed a different pairing of the letters and numbers in the compounds so that the same compound never appeared more than once. Exposure time for each stimulus in the test series was five seconds, during which time the response was written.

Five experimental groups are distinguished on the basis of the relative amount of training offered on letters and numbers. For Group I of 30 subjects, two exposures of each letter pair and each number pair were provided in each learning series (L:N = 2:2); for Group II of 28 subjects (L:N = 2:1); for Group III of 33 subjects (L:N = 1:2); for Group IV of 26 subjects (L:N = 3:1) and for Group V of 28 subjects (L:N = 1:3).

The measure of the probability of correct response to the letters, the numbers, and the four classes of the compound is obtained for each of the 15 test series by pooling over stimuli and subjects. Table 2 contains these data.

TABLE 2

Obtained Probability of Correct Response for Letters Alone, Numbers Alone and the Four Classes of the Letter-Number Compounds

Group	I (L:N=2:2)			II (L:N=2:1)			III (L:N=1:2)			IV (L:N=3:1)			V (L:N=1:3)		
Test	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
$P\{C\mid L\}$	.59	.67	.74	.57	.62	.62	.55	.63	.67	.66	.76	.80	.51	.59	.62
$P\{C\mid N\}$	.56	.73	.78	.53	.62	.64	.56	.71	.74	.56	.63	.70	.59	.68	.82
$\Pr\{C\mid(L_C, N_C)\} \cdot \Pr\{L_C, N_C\}$	.25	.46	.57	.24	.33	.33	.22	.41	.48	.30	.44	.55	.22	.38	.48
$\Pr\{C\mid(L_C, N_I)\} \cdot \Pr\{L_C, N_I\}$	.13	.10	.10	.12	.12	.14	.11	.09	.07	.23	.20	.17	.11	.03	.05
$\Pr\{C\mid(L_I, N_C)\} \cdot \Pr\{L_I, N_C\}$	.13	.17	.15	.13	.12	.14	.18	.20	.17	.09	.07	.09	.22	.23	.25
$\Pr\{C\mid(L_I, N_I)\} \cdot \Pr\{L_I, N_I\}$	.08	.02	.02	.11	.07	.06	.08	.04	.04	.04	.05	.02	.09	.04	.02
$\Pr\{C\mid(L, N)\}$	.59	.75	.84	.60	.64	.67	.59	.74	.76	.66	.76	.83	.64	.68	.80

A first test of the model comes from the relationship $\Pr\{C|L\}\cdot\Pr\{C|N\} - \Pr\{C|(L_C, N_C)\}\cdot\Pr\{L_C, N_C\} = \Pr\{C|(L_I, N_I)\}\cdot\Pr\{L_I, N_I\}$ derived from (10) and (13). The differences between the values for the left side and the right side of the equations for the 15 observations have a mean of $-.004$, a range of $-.03$ to .05, and a root mean square deviation around 0 of .02. It would appear that the fit is sufficiently good so that (11), (12), and (15) may be used in a further test of the model.

TABLE 3

Derived Values for $\Pr\{D_L \geq d\}$ and $\Pr\{D_N \geq d\}$ and Predicted Probability of Correct Response for the Four Classes of the Letter-Number Compounds

Group	I (L:N=2:2)			II (L:N=2:1)			III (L:N=1:2)			IV (L:N=3:1)			V (L:N=1:3)		
Test	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
$\Pr\{D_L \geq d\}$	.27	.43	.55	.25	.37	.33	.08	.30	.40	.41	.58	.67	-.09	.19	.28
$\Pr\{D_N \geq d\}$	.27	.57	.65	.27	.37	.31	.22	.53	.62	.13	.30	.52	.13	.62	.69
$\Pr\{C\|(L_C, N_C)\} .\Pr\{L_C, N_C\}$	.28	.47	.56	.26	.35	.35	.23	.42	.48	.32	.45	.55	.16	.39	.49
$\Pr\{C\|(L_C, N_I)\} .\Pr\{L_C, N_I\}$	.14	.11	.11	.13	.15	.16	.11	.10	.09	.23	.23	.17	.10	.05	.07
$\Pr\{C\|(L_I, N_C)\} .\Pr\{L_I, N_C\}$	.14	.18	.16	.14	.15	.15	.18	.22	.20	.09	.09	.10	.21	.26	.27
$\Pr\{C\|(L_I, N_I)\} .\Pr\{L_I, N_I\}$	.05	.02	.10	.04	.03	.05	.08	.03	.02	.05	.03	.01	.14	.01	.02
$\Pr\{C\|L, N\}$	.61	.78	.84	.57	.68	.71	.60	.77	.79	.69	.80	.83	.61	.71	.85

With three equations, the two unknowns are overdetermined, and since there is no a priori reason for selecting any particular pair of equations for solution, it was decided to obtain all three solutions and use the means as the best estimates of Pr $\{D_L \geq d\}$ and Pr $\{D_N \geq d\}$. Accordingly, the appropriate empirical values for each of the three tests for each of the five groups were substituted into the equations and solutions for Pr $\{D_L \geq d\}$ and Pr $\{D_N \geq d\}$ obtained. The resulting mean values are shown in Table 3. To obtain an indication of the consistency of the three equations, the standard deviation of the three estimates of Pr $\{D_L \geq d\}$ and Pr $\{D_N \geq d\}$ for each of the 15 determinations was computed. These values ranged from .01 to .19 and yielded a mean and median of .09 and .09, respectively, for Pr $\{D_L \geq d\}$. The range for Pr $\{D_N \geq d\}$ was .01 to .19 with a mean and median of .10 and .12, respectively.

The predicted values for Pr $\{C|(L_C, N_C)\} \cdot$ Pr $\{L_C, N_C\}$, Pr $\{C|(L_C, N_I)\} \cdot$ Pr $\{L_C, N_I\}$, Pr $\{C|(L_I, N_C)\} \cdot$ Pr $\{L_I, N_C\}$, Pr $\{C|(L_I, N_I)\} \cdot$ Pr $\{L_I, N_I\}$, and Pr $\{C|L, I\}$ using mean Pr $\{D_L \geq d\}$ and mean Pr $\{D_N \geq d\}$ are contained in Table 3. A comparison of the obtained values of Table 2 with the predicted values of Table 3 shows a quite satisfactory fit except for a small consistent tendency for the predicted values of Pr $\{C|(L_C, N_C)\} \cdot$ Pr $\{L_C, N_C\}$, Pr $\{C|(L_C, N_I)\} \cdot$ Pr $\{L_C, N_I\}$, and Pr $\{C|(L_I, N_C)\} \cdot$ Pr $\{L_I, N_C\}$ to be too high and the predicted values of Pr $\{C|(L_I, N_I)\} \cdot$ Pr $\{L_I, N_I\}$ to be too low. This discrepancy is reflected in mean differences of .007, .012, .013, and $-.013$ between predicted and obtained values for these classes of the compound stimuli. The root mean square deviations of the differences around 0 for the same four classes of the compound stimuli and for the total, Pr $\{C|L, N\}$, yielded values of .021, .016, .018, .028, and .031, indicating further the adequacy of the fit of the model to these data.

REFERENCES

[1] Bush, R. R. and Mosteller, F. A mathematical model for simple learning. *Psychol. Rev.*, 1951, **58**, 313-323.

[2] Estes, W. K. and Burke, C. J. A theory of stimulus variability in learning. *Psychol. Rev.*, 1953, **60**, 276-286.

[3] Gulliksen, H. A generalization of Thurstone's learning function. *Psychometrika*, 1953, **18**, 297-307.

[4] Hull, C. L. Principles of behavior. New York: Appleton-Century-Crofts, 1943.

[5] Restle, F. A theory of discrimination learning. *Psychol. Rev.*, 1955, **62**, 11-19.

[6] Schoeffler, M. O. Probability of response to compounds of discriminated stimuli. *J. exp. Psychol.*, 1954, **48**, 323-329.

[7] Spence, K. W. The nature of discrimination learning in animals. *Psychol. Rev.*, 1936, **43**, 427-449.

[8] Thurstone, L. L. The learning function. *J. gen. Psychol.*, 1930, **3**, 469-493.

Manuscript received 12/10/56

Revised manuscript received 2/20/57

PSYCHOMETRIKA—VOL. 22, NO. 4
DECEMBER, 1957

PROCEDURES FOR OBTAINING SEPARATE SET AND CONTENT COMPONENTS OF A TEST SCORE*

GERALD C. HELMSTADTER

COLORADO STATE UNIVERSITY†

Using two distinct models, several formulas for obtaining separate set and content components of a test score have been derived. Comparisons among the methods are made algebraically and through their application to a set of test data apparently affected by response sets.

Cronbach [1] has pointed out that when a test is composed of difficult items having but two or three alternatives, a response set is likely to affect the total test score. By response set is meant "any tendency causing a person consistently to give different responses to test items than he would when the same content is presented in different form." Cronbach has further shown [2] that such an effect shows test-retest stability, and he feels that there is adequate evidence to conclude that various response sets reflect "real" dimensions of human differences. As he further points out, some of the response-set variance is potentially useful while some of it will interfere with measurement. To be able to capitalize on the effect of response set when it is useful and to eliminate it when it is undesirable, some procedure for obtaining separate set and content components of a test score is necessary. This paper presents a logical basis and compares a number of scoring procedures for separating the response set from the score reflecting individual differences with respect to the obvious item content. Because response sets most commonly occur in tests composed of items which have but two alternatives, the following discussion is restricted to this case. Also, for convenience, it will be assumed that no items are omitted.

To simplify the discussion a single set of notation will be used throughout. For convenience, these are listed together below. Further explanation of the terms will be made as each is used.

*The author wishes to thank Dr. Norman Frederiksen, who brought this problem to his attention, and Drs. Ledyard R Tucker and Frederic M. Lord, who suggested three of the procedures. All have contributed materially to this paper through many helpful suggestions and criticisms.

†This paper was written while the author was an Associate in Research at the Educational Testing Service, Princeton, New Jersey.

N_α and N_β = number of items keyed A and B, respectively.*

K_α and K_β = number of items keyed A and B, respectively, which have been answered correctly on the basis of content.

P_A and P_B = the examinee's probability of marking an item not answered on the basis of content A and B, respectively.

R_A = number of items keyed A and marked A by the examinee.

R_B = number of items keyed B and marked B by the examinee.

W_A = number of items keyed B but marked A by the examinee.

W_B = number of items keyed A but marked B by the examinee.

$$\left.\begin{matrix} A_{\alpha a_i}\,, & A_{\beta a_i} \\ A_{\alpha b_i} & A_{\beta b_i} \end{matrix}\right\} = \left\{\begin{matrix}\text{the proportion of examinees who marked } A \text{ for an item classified by the} \\ \text{subscript as follows: } \alpha = \text{keyed } A;\ \beta = \text{keyed } B;\ a_i = \text{marked } A \text{ by} \\ \text{person } i;\ b_i = \text{marked } B \text{ by person } i.\end{matrix}\right.$$

C_j = content score by procedure j.

S_j = set score by procedure j. (S_j will always mean a set to mark response A.)

The Scoring Procedures

Usual Total Score Procedure

Ordinarily, a test is scored by simply counting the number of items in agreement with a key. Thus, in the present terminology, the usual scoring procedure can be expressed as

$$C_1 = R_A + R_B\ . \tag{1}$$

In such a procedure, no attempt is made to distinguish between the relative effects of content and set. However, a straightforward (though not entirely satisfactory) way of obtaining an estimate of set would be to take the difference between the number of items the examinee marked A and the number he marked B. Thus, let

$$S_1 = R_A + W_A - (R_B + W_B). \tag{2}$$

Warranted Set Procedure

One possible solution to the problem of eliminating the effects of a response set is to give the examinee a score based upon the extent to which his tendency to mark response A rather than response B is warranted.

Consider the relationship between the keyed response and the examinee's response illustrated in Figure 1. One could think of the elements of this matrix as follows:

R_A and R_B = examinee's warranted set to mark A and to mark B, respectively;

W_A and W_B = examinee's unwarranted set to mark A and to mark B, respectively;

*Throughout this paper it will be helpful to remember that the Greek subscripts indicate the way an item was keyed, while the English subscripts indicate the way an item was marked by the individual or by the group of examinees.

		Examinee's	Response	
		A	B	Total
Keyed Response	α	R_A	W_B	N_α
	β	W_A	R_B	N_β
	Total	$R_A + W_A$	$R_B + W_B$	$N_\alpha + N_\beta$

Fig. 1. Observed Score Matrix

$$\left.\begin{array}{c} R_A + W_A \\ \text{and} \\ R_B + W_B \end{array}\right\} = \text{examinee's total set to mark } A \text{ and to mark } B\text{, respectively.}$$

Then, if any indeterminate ratios are put equal to zero, a content score could be defined as

$$C_2 = \frac{\text{warranted set to mark } A}{\text{total set to mark } A} + \frac{\text{warranted set to mark } B}{\text{total set to mark } B}$$

$$= \frac{R_A}{R_A + W_A} + \frac{R_B}{R_B + W_B}. \tag{3}$$

Similarly, it would be possible to define a set scores as

$$S_2 = \frac{\text{unwarranted set to mark } A}{\text{total set to mark } A} - \frac{\text{unwarranted set to mark } B}{\text{total set to mark } B}$$

$$= \frac{W_A}{R_A + W_A} - \frac{W_B}{R_B + W_B}. \tag{4}$$

Postulated Knowledge Procedure

A third way which might be used to score a test which is subject to the effects of a response set involves an estimation of the number of responses actually based on content. Following a logic somewhat similar to that used in deriving scores which "correct for guessing," the number of items observed as marked in agreement with the key (i.e., R_A and R_B) can be thought of as resulting from a summation of those items marked in a given direction on the basis of content and of those items marked in this same direction on the basis of set. If K_α and K_β represent the number of α-items and β-items marked A and B, respectively, on the basis of content, then $N_\alpha - K_\alpha$ and $N_\beta - K_\beta$ represent the number of α-items and β-items, respectively, which have been answered on some other basis. Then, if P_A and P_B represent the probability that an individual marks an item A or B, respectively, on a basis other than content, the number of α-items and the number of β-items marked in agreement with the key on a basis other than content will be $P_A(N_\alpha - K_\alpha)$ and $P_B(N_\beta - K_\beta)$, respectively. Thus, it can be said that

$$R_A = K_\alpha + P_A(N_\alpha - K_\alpha) \tag{5}$$

and

$$R_B = K_\beta + P_B(N_\beta - K_\beta). \tag{6}$$

If no omits are permitted, the examinee must mark either A or B and thus

$$P_A + P_B = 1. \tag{7}$$

A final equation necessary to obtain values for K_α , K_β , P_A , and P_B can be obtained by assuming that the α-items are equal in difficulty to the β-items. This assumption can be expressed as

$$\frac{K_\alpha}{N_\alpha} = \frac{K_\beta}{N_\beta}. \tag{8}$$

The simultaneous solution of these last four equations yields

$$K_\alpha = R_B\left(\frac{N_\alpha}{N_\beta}\right) - W_B , \tag{9}$$

$$K_\beta = R_A\left(\frac{N_\beta}{N_\alpha}\right) - W_A , \tag{10}$$

$$P_A = \frac{N_\alpha W_A}{N_\beta W_B + N_\alpha W_A} , \tag{11}$$

$$P_B = \frac{N_\beta W_B}{N_\beta W_B + N_\alpha W_A}. \tag{12}$$

Given these values, the obvious content score is

$$C_3 = K_\alpha + K_\beta \tag{13}$$

$$= R_A\left(\frac{N_\beta}{N_\alpha}\right) + R_B\left(\frac{N_\alpha}{N_\beta}\right) - (W_A + W_B). \tag{14}$$

While P_A itself could serve as a measure of set toward A, it is more convenient to let

$$S_3 = 2P_A - 1. \tag{15}$$

Here S_3 ranges from -1 to $+1$ and is 0 when the examinee is just as likely to mark an unknown item A as he is to mark it B. Thus,

$$S_3 = \frac{2N_\alpha W_A}{N_\beta W_B + N_\alpha W_A} - 1. \tag{16}$$

Orthogonal Score Procedure

Another possibility is to consider the set and content scores as orthogonal traits. In this conception of the problem, each examinee is represented by a point in a plot of the proportion of items keyed A which were marked A (i.e., R_A/N_α) against the proportion of items keyed B which were marked A. If this is done, the vector going from (0, 0) to (1, 1) can be considered a set axis and the vector going from (0, 1) to (1, 0) a content axis. The set and content scores of the examinee can then be defined as some function of the projection of his plotted point on these respective axes, e.g., that given in Figure 2.

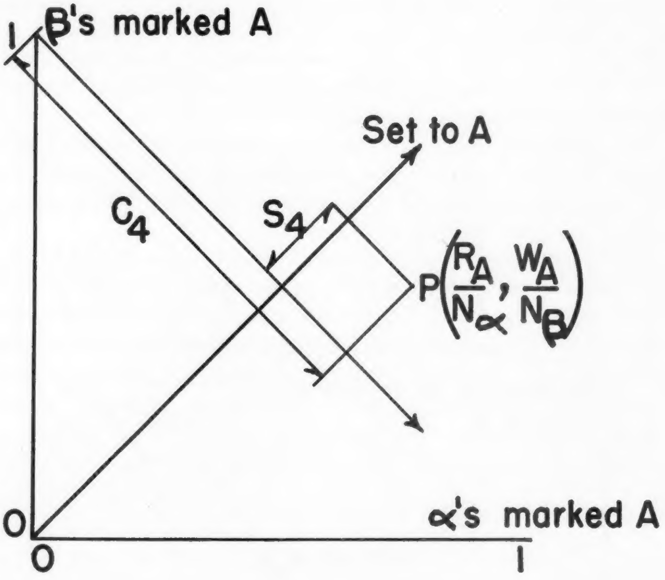

Fig. 2 Set and Content as Orthogonal Traits.

The scores can readily be obtained in terms of the observed measures by applying a 45° rotation and making an appropriate translation of the axes. For convenience, translations which make all content scores positive and which make set scores equal to zero when at a chance level (i.e., when $P_A = P_B = \frac{1}{2}$) have been used. Thus, by this procedure:

$$C_4 = .707\left(\frac{R_A}{N_\alpha} - \frac{W_A}{N_\beta} + 1\right); \tag{17}$$

$$S_A = .707\left(\frac{R_A}{N_\alpha} + \frac{W_A}{N_\beta} - 1\right). \tag{18}$$

Postulated Scale Score

All of the procedures thus far discussed have assumed that the items in the test were dichotomous with respect to content. That is, the items have been considered to represent either A or not A. The solutions to the problem of obtaining separate set and content scores presented in this and the following sections make a different assumption: that the extent to which items represent A can be expressed as a continuous variable. Thus, the test items are visualized as falling along a unidimensional scale characterized by the content. For example, in an inventory designed to detect an authoritarian personality, it might be preferable to think of statements (with which a respondent is asked to indicate his agreement or disagreement) as representing various degrees of authoritarianism rather than as being classified as authoritarian and nonauthoritarian.

One possibility under this view is to postulate, for each individual, a characteristic curve relating the probability of his marking an item of a given scale value as A to the scale value of the item. Two such curves are illustrated in Figure 3.

The content score of the individual would be his ability to distinguish between items having different scale values and thus would be represented by the slope of the curve. On the other hand, the set score measures his tendency to call all items A or all items B and therefore would be represented

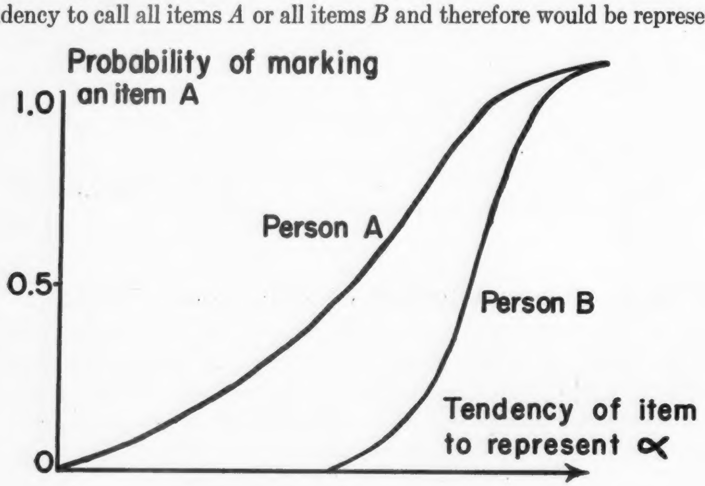

Fig. 3. Item Characteristic Curve for Two Individuals.

by some index of central tendency which would locate the position of the curve on the scale. The scale value corresponding to probability 1/2 of marking an item A could be used for this purpose.

The problem, then, is to determine the important parameters of this characteristic curve from observations of one or zero (i.e., calling an item A or not calling an item A) for each person for each item. While theoretically it would be possible to obtain the maximum likelihood estimators of the desired parameters, preliminary work, specifying first the normal ogive and then a straight line as the form of the characteristic curve, indicated that the solutions are far too complex to be of practical value.

One approximation which is feasible, however, is the following. Assume that all the items fall at only two points which differ along the scale characterized by A. Those items keyed A could be considered as estimates of one point, and those keyed B of the second. Then the scale value of each of these points could be obtained by averaging, within each group of items, the normal deviates corresponding to the proportion of persons marking the item A, that is by taking

$$\bar{Z}_\alpha = \frac{1}{N_\alpha} [\sum_{a_i} Z_{A\alpha a_i} + \sum_{b_i} Z_{A\alpha b_i}], \tag{19}$$

and

$$\bar{Z}_\beta = \frac{1}{N_\beta} [\sum_{a_i} Z_{A\beta a_i} + \sum_{b_i} Z_{A\beta b_i}], \tag{20}$$

where Z is the normal deviate corresponding to the indicated proportions. Then, the normal deviates for the proportion of A items and the proportion of B items which *each individual* marked A can be plotted at these points as indicated in Figure 4. The slope of the line determined by these two points can now be used as a content score, and the height of the line at its midpoint can be used as a set score. Thus,

$$C_5 = \frac{Z_{R_A/N_\alpha} - Z_{W_A/N_\beta}}{\bar{Z}_\alpha - \bar{Z}_\beta} \tag{21}$$

and

$$S_5 = \tfrac{1}{2}(Z_{R_A/N_\alpha} + Z_{W_A/N_\beta}). \tag{22}$$

Correlation Procedure

An alternative procedure for obtaining a content score, assuming the items can be scaled, is to compute the biserial correlation between the examinees' dichotomized responses and scale values of the items along a content continuum. Both Tucker [6] and Lord [4] have indicated that this biserial correlation is a simple function of the slope of the characteristic curve whenever the scale values of the items have a normal distribution for the particular

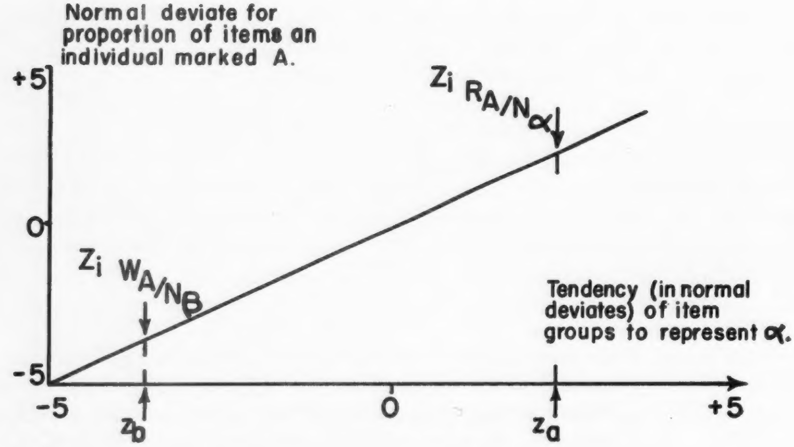

Fig. 4 . An Approximation to the Item Characteristic Curve.

test under consideration. Thus, the correlation procedure provides another means of estimating the slope of the characteristic curve shown in Figure 3. Also, it seems quite reasonable to consider an examinee who can successfully rank all the items in the test according to their relative position along the content continuum as having the ability to make very good discriminations with respect to the content, regardless of where he would locate the items as a group along the axis.

The formula for biserial correlation is

$$r_b = (M_2 - M_1)pq/Z\sigma_y \,. \tag{23}$$

If the proportion of examinees (in a standard group) that marked the item A is used as the scale value of the item, the elements for the biserial correlation formula can be expressed in terms of the notation used here as follows:

$$p = \frac{R_A + W_A}{N_\alpha + N_\beta}\,; \qquad q = \frac{R_B + W_B}{N_\alpha + N_\beta}\,;$$

$$M_1 = \frac{\sum_\alpha A_{\alpha b_i} + \sum_\beta A_{\beta b_i}}{R_B + N_\alpha - R_A}\,; \qquad M_2 = \frac{\sum_\alpha A_{\alpha a_i} + \sum_\beta A_{\beta a_i}}{R_A + N_\beta - R_B}\,;$$

$\sigma_y = \sigma_A$; $r_b = C_6$; Z = normal deviate corresponding to p.
There seems to be no direct suggestion for a set score by a correlational procedure.

Comparison of the Methods

When several approaches are proposed for the solution of a single problem, it is imperative that some attempt be made to determine the extent to which the various solutions produce similar results. Thus, all of the formulas have been compared and those which were not algebraically identical nor linearly equivalent used to obtain separate set and content scores for 62 individuals who had been given an experimental test designed to measure one aspect of report-writing ability.

First consider the content scores, writing them solely in terms of the readily obtained quantities R_A , R_B , N_α , and N_β .

$$C_1 = R_A + R_B \; ; \tag{24}$$

$$C_2 = \frac{R_A}{N_\beta + (R_A - R_B)} + \frac{R_B}{N_\alpha - (R_A - R_B)} \; ; \tag{25}$$

$$C_3 = R_A(N_\beta/N_\alpha) + R_B(N_\alpha/N_\beta) - (N_\beta - R_B + N_\alpha - R_A) \tag{26}$$

$$= (N_\alpha + N_\beta)[(R_A/N_\alpha) + (R_B/N_\beta) - 1] \tag{27}$$

$$C_4 = .707\{(R_A/N_\alpha) - [(N_\beta - R_B)/N_\beta] + 1\} \tag{28}$$

$$= .707[(R_A/N_\alpha) - (R_B/N_\beta)] \tag{29}$$

$$= \frac{.707}{N_\alpha + N_\beta} C_3 + .707; \tag{30}$$

$$C_5 = \frac{Z_{R_A/N_\alpha} - Z_{(N_\beta - R_A)/N_\beta}}{\bar{Z}_\alpha - \bar{Z}_\beta}. \tag{31}$$

If $\bar{Z}_\alpha - \bar{Z}_\beta$ be the unit of the scale, noting that $Z_{1-p} = -Z_p$,

$$C_5 = Z_{R_A/N_\alpha} + Z_{R_B/N_\beta} \, . \tag{32}$$

$$C_6 = \frac{[N_\alpha - (R_A - R_B)][\sum_\alpha A_{\alpha a_i} + \sum_\beta A_{\beta a_i}]}{Z\sigma_A(N_\alpha + N_\beta)^2} - \frac{[N_\beta + (R_A - R_B)][\sum_\alpha A_{\alpha b_i} + \sum_\beta A_{\beta b_i}]}{Z\sigma_A(N_\alpha + N_\beta)^2} \tag{33}$$

$$= \frac{p(\sum_\alpha A_{\alpha a_i} + \sum_\beta A_{\beta a_i}) - q(\sum_\alpha A_{\alpha b_i} + \sum_\beta A_{\beta b_i})}{Z\sigma_A(N_\alpha + N_\beta)}, \tag{34}$$

where

$$p = \frac{N_\alpha - (R_A - R_B)}{N_\alpha + N_\beta} \quad \text{and} \quad q = \frac{N_\beta + (R_A - R_B)}{N_\alpha + N_\beta}.$$

Expressed in these terms it is readily apparent that C_3 and C_4 will place examinees in the same order and are linear functions of $(R_A/N_\alpha) + (R_B/N_\beta)$ which, for convenience, will be called the simplified ratio score and be designated by C_7 . Also, it is interesting to note that when $N_\alpha = N_\beta$, C_7 is perfectly correlated with the number correct.

Next, consider the set scores, again writing each in terms of the quantities R_A , R_B , N_α, and N_β :

$$S_1 = R_A + N_\beta - R_B - R_B - N_\alpha + R_A \tag{35}$$

$$= N_\beta - N_\alpha + 2(R_A - R_B); \tag{36}$$

$$S_2 = \frac{N_\beta - R_B}{R_A + N_\beta - R_B} - \frac{N_\alpha - R_A}{R_B + N_\alpha - R_A} \tag{37}$$

$$= \frac{R_B}{N_\alpha - (R_A - R_B)} - \frac{R_A}{N_\beta + (R_A - R_B)} ; \tag{38}$$

$$S_3 = \frac{2N_\alpha(N_\beta - R_B)}{N_\beta(N_\alpha - R_A) + N_\alpha(N_\beta - R_B)} - 1 \tag{39}$$

$$= \frac{(R_A/N_\alpha) - (R_B/N_\beta)}{2 - [(R_A/N_\alpha) + (R_B/N_\beta)]} ; \tag{40}$$

$$S_4 = .707\{[(R_A/N_\alpha) + (N_\beta - R_B)/N_\beta] - 1\} \tag{41}$$

$$= .707[(R_A/N_\alpha) - (R_B/N_\beta)]; \tag{42}$$

$$S_5 = .5(Z_{R_A/N_\alpha} + Z_{(N_\beta - R_B)/N_\beta}) \tag{43}$$

$$= 5(Z_{R_A/N_\alpha} - Z_{R_B/N_\beta}). \tag{44}$$

In this instance, no two procedures will produce identical results insofar as the ranking of the examinees is concerned. Since, however, S_4 will rank individuals the same as $(R_A/N_\alpha) - (R_B/N_\beta)$, this procedure will hereafter be designated S_7 to indicate that it is the companion of the simplified ratio score, C_7 .

The examination used in the empirical comparison of the methods was designed to measure a person's ability to recognize when one expression could be substituted for another without altering the meaning of the statement. Each item consisted of a short statement containing a word or expression which had been underlined and an alternative expression in parenthesis at the end of the sentence. The task was to indicate whether or not the alternative expression could be substituted for the original one without influencing the possible consequences should some policy decision, administrative action, or legal claim hinge upon the interpretation of the statement. In scoring the test, 70 such statements were used. Forty-five items were keyed same (i.e., the consequences would be the same no matter which alternative expression was used) and 25 were keyed different.

In an experimental tryout, the test was administered to 62 students in a graduate school of journalism. When the results were analyzed, it was noted that while the corrected odd-even reliability of the total score was only .47, similar reliabilities, obtained by scoring separately those items keyed "same" and those keyed "different," were .79 and .72, respectively. This fact, emphasized by the resulting correlation of $-.54$ between scores obtained on the "same" items and those obtained on the "different" items, led to the conclusion that a response set was affecting the results.

Because these results suggested a real difference among individuals on a variable other than that measured by the total score, it was felt that the data would be appropriate for a comparison of the various procedures suggested for obtaining separate content and set components of a test score. Consequently C_1 and S_1 (total score and simple difference), C_2 and S_2 (content and set by the warranted set procedure), S_3 (set by the postulated knowledge procedure), C_5 and S_5 (content and set by the postulated scale score procedure) C_6 (content by correlation procedure), C_7 and S_7 (content and set by the simplified ratio procedure) were all computed from the data and the intercorrelations obtained. The results are presented in Table 1. The values below the diagonal represent the median value of the correlations in the respective block.

Discussion and Conclusion

It will be recalled that two fundamentally different models have been used in the derivation of the various indices. One model assumes that there are right and wrong answers to the items and that the degree of set can be determined from the answers which disagree with a key. The total score, the warranted set, the simplified ratio and the postulated knowledge procedures are definitely of this type. The other model, of which the prototype is the correlation procedure, assumes instead that the items can be scaled along a continuum. The postulated scale score, while based on the latter model, requires the use of a key to obtain a practical solution, and thus might be considered a compromise.

In view of both a comparison of the formulas and the results of the empirical illustration presented, it is tempting to conclude that in many instances it will make little difference which method is used. This is particularly true with respect to measures of set where the intercorrelations of the methods vary from .94 to .99. However, McCornack [5] has pointed out the danger of assuming that just because two keys correlate highly they will have the same external validity. He shows, for example, that given two keys which correlate .94 with one another, when one correlates .60 with an external

TABLE I

Intercorrelation Among Various Set and Content Components of a Test Score*

N = 62

	Content Scores					Set Scores				
Procedure	Total Score	Warranted Set	Postulated Scale	Correlation	Simplified Ratio	Simple Difference	Warranted Set	Postulated Knowledge	Postulated Scale	Simplified Ratio
	C_1	C_2	C_5	C_6	C_7	S_1	S_2	S_3	S_5	S_7
C_1		.92	.71	.75	.88	.53	.57	.30	.32	.39
C_2			.88	.78	.92	.03	.38	.07	.07	.15
C_5		.82		.70	.87	-.07	-.01	-.33	-.34	-.21
C_6					.78	.17	.22	-.03	-.01	.05
C_7						.05	.11	-.18	-.17	-.10
S_1							.97	.94	.95	.99
S_2								.94	.94	.95
S_3			.05						.98	.97
S_5							.96			.97
S_7										

*For 60 degrees of freedom, the 5% and 1% values of r are .250 and .325 respectively.

criterion the other might correlate anywhere between .29 and .84 with that same criterion. It is extremely important, therefore, to make use, if it is at all feasible, of an external criterion in selecting a method for obtaining separate set and content components of a test score.

When no external criterion is available, other considerations become important. Thus, for example, it might be noted that in the present illustration the simplified ratio procedure has the highest first centroid loading in the matrix of set scores, has the next to the highest loading in the matrix of content scores, and has an average correlation between content and set scores which is closer to zero than that for any other method. That this occurred in a case which would clearly be more appropriate for the continuity model is particularly encouraging since use of the continuity model requires considerable extra work in the scaling of the items and in the computation of the scores.

Obviously, such evidence as that presented here does not conclusively establish the efficacy of the content and set scores described in this paper. To do this it would be necessary to carry out studies which indicated whether or not the set can be independently manipulated through the use of various experimental controls. For example, one might design an experiment which would involve the administration of a test such as that of alternative expressions under two or more conditions that differ in the extent to which set is likely to occur. Or, one might try writing tests which would yield the same content scores but different set scores when different item forms were used. Beyond this, the usefulness of such scores would have to be established by the usual reliability and validity studies in the context of a particular applied situation.

This evidence does suggest, however, that one or more of the procedures developed here might be useful in a number of situations. In the absence of an external criterion against which to compare the methods and without further experimental evidence, present indications are that the simplified ratio procedure will provide an adequate approximation to set and content components of a test score except when both extreme accuracy is needed and when, in addition, the continuity model is obviously the most appropriate. In this latter case, use of the correlation procedure to obtain content scores would appear to be justified.

REFERENCES

[1] Cronbach, L. J. Responses sets and test validity. *Educ. psychol. Measmt*, 1946, 6, 475-494.

[2] Cronbach, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt*, 1950, 19, 3-31.

[3] Gage, N. L. and Cronbach, L. J. Conceptual and methodological problems in interpersonal perception. *Psychol. Rev.*, 1955, 62, 401-422.

[4] Lord, F. M. A theory of test scores. *Psychometric Monogr. No. 7*, 1952. Chicago: Univ. of Chicago Press.

[5] McCornack, R. L. A criticism of studies comparing item-weighting methods. *J. appl. Psychol.*, 1956, 40, 343-344.

[6] Tucker, L. R. Maximum validity of a test with equivalent items. *Psychometrika*, 1946, 11, 1-13.

Manuscript received 11/29/56

Revised manuscript received 4/11/57

PSYCHOMETRIKA—VOL. 22, NO. 4
DECEMBER, 1957

ITEM SELECTION METHODS FOR INCREASING TEST HOMOGENEITY

HAROLD WEBSTER

VASSAR COLLEGE

A number of methods for increasing test homogeneity by item selection are discussed. Exact selection conditions which will maximize obtained homogeneity as measured by KR - 20 and KR - 21 are derived, and an application is given. Since they require only item count data, the selection conditions are economical to apply.

A problem which is likely to arise whenever psychological tests are constructed is that of adding or discarding items in such a way that the resulting test will have some optimum degree of homogeneity or "split-test" reliability. Adkins [1] and Davis [4] have reviewed various practical solutions in general use. A popular method consists in retaining items with high item-test correlations and discarding those with low correlations, but this is not the best way to increase obtained homogeneity as measured by the reliability coefficients in common use, KR-20 and KR-21, originally derived by Kuder and Richardson [10]. The purpose of the present paper is to present exact and economical item selection methods for increasing Kuder-Richardson reliability.

There has been some debate concerning the adequacy of KR-type coefficients as measures of test homogeneity [11, 9]. In this paper homogeneity will be defined either as KR-20 or else as a rather general coefficient due to Lord [13], which is formally the same as KR-21; which coefficient is better to use in the item selection conditions probably depends, to judge from applications, on the sampling theory presented by Lord [14]. No attempt will be made to improve these definitions of homogeneity even though inadequacies are recognized, some of which are mentioned in the next paragraph.

Although the item selection conditions which will be derived could be used to *maximize* homogeneity, there are important reasons why maximizing homogeneity for a given sample will usually be impractical. First, for a given sample small increases in homogeneity, especially in its upper range, are likely to be lost in subsequent samples because of unknown sampling variations. Second, although cases where homogeneity is impractically high are seldom encountered, it is known (when items are assigned point scores) that tests with homogeneity approaching unity will have undesirable item redundancy [7, 16]. Finally, when using items for which the direction of scoring is fixed, increasing homogeneity beyond a certain value tends rapidly to increase the proportion of items retained which have extreme means, which in turn increases skewness of the test distribution. For these reasons, homogeneity somewhat less than the maximum seems practical.

There are unsolved statistical problems, including the sampling behavior of KR-20 and KR-21, which will not be considered here; these problems are especially serious if it is necessary to increase homogeneity for a number of subtests simultaneously. Investigators who are primarily interested in obtaining group-factor subtests from numerous items have available the methods of Wherry and Winer [20] and of Loevinger, Gleser and DuBois [12]. The present paper considers the more limited problem of increasing by item selection the observed homogeneity of a single test.

Gulliksen ([8], p. 379) has suggested a graphical method for selecting for retention in a test those items which increase homogeneity as measured by KR-20. Discarded items are those for which the ratio of item variance to reliability index (item-test covariance divided by test standard deviation) is relatively large. Iterative conditions are derived in the present paper which use KR-20 or KR-21, which allow re-examination later of previously rejected items to see if they should be put back in the test, and which do not require plotting.

In some problems the level of precision could be increased slightly if interitem relationships were utilized directly in assessing increases in homogeneity, but usually such a refinement does not justify the additional computational labor required. All methods discussed in the present paper require, in addition to item means and variances, only item-test relationships and consequently they may be applied using only item count data.

Item Selection Conditions Which Are Independent of Test Length

In this and the next section several item selection methods for increasing test homogeneity, including the popular one mentioned in the first paragraph, will be considered. Methods which are independent of test length are discussed first, since it can be shown that they do not always result in increased homogeneity. In the next section some item selection conditions are derived which are dependent on test length and which necessarily increase KR-20 and KR-21.

The usual definition of reliability is

$$(1) \qquad r_{TT} = (V_T - E_T)/V_T = 1 - (E_T/V_T),$$

where V_T is the total variance and E_T is the error variance of test T. Either KR-20 or KR-21, which are used to define homogeneity, may be obtained directly from (1), depending upon how E_T is defined. In addition to selecting (or discarding) items in such a way that the entire ratio (1) is increased, there are other ways in which r_{TT} might be increased: by selecting those items which (*i*) increase the true variance $V_T - E_T$, or (*ii*) decrease E_T, or (*iii*) increase V_T. Conditions derived to achieve any one of these aims alone appear to have serious limitations when used in iterations for the purpose of increasing r_{TT}. For example, it was found in several applications that no single item existed which if discarded or added would decrease E_T. Also, methods (*i*) and (*iii*) are inefficient; because of the relative stability of E_T, method (*i*) appears to have much the same disadvantages as (*iii*), the inefficiency of which will next be shown.

The variance of test T minus item j may be written

$$V_{T-j} = V_T + V_j - 2C_{jT}, \tag{2}$$

where C_{jT} is the covariance of item j and test T. If item j satisfies the condition $V_T < V_{T-j}$, then it could be discarded to increase the test variance, and this is seen, using (2), to be the same as discarding j if

$$C_{jT} < V_j/2. \tag{3}$$

Dividing both sides of (3) by the product of standard deviations, S_jS_T, j is discarded if

$$r_{jT} < S_j/2S_T. \tag{4}$$

By a similar development *adding* an item k not already in T will increase the variance if

$$r_{kT} > -S_k/2S_T. \tag{5}$$

Now item selection based on alternate applications of (4) and (5) will always increase test variance. The quantities on the right in (4) and (5) are, however, quite small, and as the standard deviation S_T increases, these expressions approach even more closely the condition that items be discarded or added merely if their test correlations are, respectively, negative or positive. But the latter condition is known to be an inefficient method for increasing homogeneity, for it can be seen that one effect of applying (4) and (5), if enough items are available, will be to form a very long test. It has also been shown by other methods that if a test is long enough, practically all items with item-test correlations exceeding zero will, if added to the test, contribute to its homogeneity [3, 17]. Items with low correlations may contribute very little, however, and efficiency in testing requires that the shortest tests which achieve a specified homogeneity be used. Bedell [2] derived equations which could be solved for the number of items with lowest item-test correlations to be discarded in order to maximize the reliability of a single-factor test. Unit rank for the item matrix was assumed, and some computational approximations were developed.

A popular method is to discard item j from test T if

$$r_{jT} < k, \tag{6}$$

where k is a positive constant. The requirement that retained items be significantly correlated with their own test is approximately satisfied if a number of items satisfying (6) are discarded at once when k is some multiple of the standard error of r_{jT}. But if moderately large samples of subjects and items are available, and k is chosen to correspond to one of the usual levels of significance, then the obtained homogeneity is usually found to be decreased after applying (6).

Another method which is independent of test length will next be discussed briefly. If item j satisfies the inequality

$$r_{TT} < r_{(T-j)(T-j)}, \tag{7}$$

where the r's are homogeneity coefficients for test T and for test T *minus item j*, respectively, then discarding j will increase homogeneity. A similar expression can be written for the case where adding an item to T increases homogeneity. But whether or not j satisfies (7), or the corresponding addition condition, could depend, especially for short tests, on the length of T. It might therefore be argued that if items were to remain in T on their own merits, (7) should be rewritten so that it is independent of the length of T. It can be shown, however, that this is not advantageous, for it leads to expressions the application of which will eventually decrease homogeneity. To show this, first multiply the left side of (7) by $(n - 1)/(n - r_{TT})$. This is the change required to make (7) independent of test length, a fact which can be proved by next rearranging the terms to correspond to functions which are known to be invariant with respect to test length ([8], p. 85). Finally members of this rearranged inequality can be shown (by adding 1 to both sides and taking reciprocals) to have exactly the same effect in item selection as if they were estimates of the squared average item-test correlation. Therefore if (7) is directly altered to make it independent of test length, its application will always increase the average item-test correlation, but will not always increase homogeneity. In fact if it is used in iterations, it will reject successive halves (approximately) of the items in the test, and the analogous addition condition will not admit back into the test later any items previously rejected; consequently r_{TT} decreases sharply after the test is shortened beyond a certain point.

Item Selection Conditions Which Are Dependent on Test Length

We return therefore to (7) as the condition for discarding item j in order to increase test homogeneity. As in (1), the homogeneity of the shortened test is

$$r_{(T-j)(T-j)} = 1 - (E_{T-j}/V_{T-j}). \tag{8}$$

Lord [13, 14] has shown that if E_T is defined as the mean of the estimated sampling variances (based on random samples of n items) of the N subjects' test scores T_i , then

$$E_T = (n\bar{T} - \bar{T}^2 - V_T)/(n - 1), \tag{9}$$

where $\bar{T}$ is the sample mean. Substitution of (9) in (1) provides a measure of reliability which is formally identical with KR-21 but which is actually more general than either the latter or KR-20. The error variance of the shortened test can be written, as in (9),

$$\begin{aligned} E_{T-j} &= [(n - 1)(\bar{T} - p_j) - (\bar{T} - p_j)^2 - V_{T-j}]/(n - 2) \\ &= [n\bar{T} - \bar{T}^2 - V_T - \bar{T} - p_j(n - 2\bar{T}) + 2C_{jT}]/(n - 2), \end{aligned} \tag{10}$$

where p_j is the mean of item j.

Substituting (1), (8), (2), (9), and (10) in (7) and simplifying, it is found that item j may be discarded to increase homogeneity (as measured either by Lord's formula or by KR-21) if

$$C_{jT} - k_1 V_j - k_2 p_j < k_3 , \tag{11}$$

where the constants in (11) are

$$\begin{aligned} k_1 &= (n - 2)(n\bar{T} - \bar{T}^2 - V_T)/2[(n - 2)(n\bar{T} - \bar{T}^2) + V_T], \\ k_2 &= (n - 1)(n - 2\bar{T})V_T/2[(n - 2)(n\bar{T} - \bar{T}^2) + V_T], \\ k_3 &= V_T(\bar{T}^2 - \bar{T} + V_T)/2[(n - 2)(n\bar{T} - \bar{T}^2) + V_T]. \end{aligned}$$

Suppose that item k is not in test T. By a derivation analogous to the above, item k may be *added* to T to increase homogeneity if

$$C_{kT} + K_1 V_k - K_2 p_k > K_3 , \tag{12}$$

where the constants are

$$\begin{aligned} K_1 &= n(n\bar{T} - \bar{T}^2 - V_T)/2(n^2\bar{T} - n\bar{T}^2 - V_T), \\ K_2 &= (n - 1)(n - 2\bar{T})V_T/2(n^2\bar{T} - n\bar{T}^2 - V_T), \\ K_3 &= V_T(\bar{T}^2 - \bar{T} + V_T)/2(n^2\bar{T} - n\bar{T}^2 - V_T). \end{aligned}$$

It is interesting that the item mean p remains explicit in (11) and (12). This is not so for conditions derived using KR-20. One can discard j to increase homogeneity as measured by KR-20 if

$$\frac{n}{n - 1}\left[\frac{V_T - \sum^n V_i}{V_T}\right] < \frac{n - 1}{n - 2}\left[\frac{V_{T-j} - \sum^n V_i + V_j}{V_{T-j}}\right], \tag{13}$$

which, using (2), can be simplified to

$$C_{jT} - h_1 V_j < h_2 , \tag{14}$$

the constants being

$$h_1 = \{[(n - 1)^2 + 1]V_T + n(n - 2) \sum V_i\}/2[V_T + n(n - 2) \sum V_i];$$

$$h_2 = V_T(V_T - \sum V_i)/2[V_T + n(n - 2) \sum V_i].$$

Similarly, adding item k to T will increase KR-20 if

$$C_{kT} - H_1 V_k > H_2 , \tag{15}$$

where

$$H_1 = n^2(V_T - \sum V_i)/2(n^2 \sum V_i - V_T),$$

$$H_2 = V_T(V_T - \sum V_i)/2(n^2 \sum V_i - V_T).$$

Applications of (11), (12), (14), *and* (15)

Computations of the item selection conditions to increase either KR-21 or KR-20 are not as laborious as they may first appear. By grouping the test distribution into five symmetrical categories as recommended by Flanagan [5], and obtaining item counts for the four extreme categories, calculation of the test variances and item-test covariances required in the conditions can be carried out rapidly.

It is known that if the T distribution is grouped into five categories containing percentages of scores (from high to low), 9, 19, 44, 19 and 9, which are assigned new scores, 2, 1, 0, -1 and -2, respectively, then the grouping will not only have high efficiency, but will also incorporate an adjustment for the effect of the coarse grouping on the estimate of r_{jT} , the item-test point-biserial correlation [18]. The new covariances will be, for a sample of N subjects,

$$C'_{jT} = (2e + f - g - 2h)/N = D_{jT}/N. \tag{16}$$

In (16) e, f, g, and h are frequencies of a preferred response to item j for the extreme categories for which the scores are 2, 1, -1, and -2, respectively. It can then be shown that the covariances needed in (11), (12), (14), and (15) are, to a sufficiently close approximation,

$$C_{jT} = D_{jT} \sum D_{iT}/N^2, \tag{17}$$

where the summation is over the n weighted differences, corresponding to the n items, as defined in (16). Also an estimate of the variance required for computing the constants of the item selection conditions is

$$V_T = (\sum D_{iT})^2/N^2. \tag{18}$$

As an example, Table 1 presents data which show how the item selection conditions increased homogeneity for a sample of 100 college women. The test used was the *De* scale [6] from the California Psychological Inventory,

TABLE 1

Variations in Homogeneity Due to Selecting Items so as to Increase KR-21 and KR-20

n	KR-21	KR-20	$\bar{x}$	V_T	Discard	Add
54	.562	.675	17.85	26.63	15	0
39	.677	.757	11.47	23.81	7	1
33	.708	.751	8.33	19.89	8	1
26	.725	.764	6.15	15.52	-	-
39	.677	.757	11.47	23.81	7	0
32	.685	.770	9.38	19.71	5	4
31	.693	.780	8.46	18.66	-	-

a personality test known to discriminate delinquent from nondelinquent persons in numerous samples. In a number of previous samples, KR-21 for the complete scale of 54 items has been found to fall in the range .50 to .60.

The first four rows of Table 1 show how KR-20 and KR-21 varied when items were alternately discarded and added back in iterations using conditions (11) and (12). The last three rows of Table 1 show variations in these same coefficients when (14) and (15) were applied starting with the 39-item test of the second row.

Discussion

The method is not very time consuming, and with the help of (17) and (18) can easily be applied using item count data. Since there were only 100 papers, the data of Table 1 required the time of one person for two days; however, this sample was used only for this example, and a much larger sample would ordinarily be preferable. It is likely that when there is only a single sample available, the greater the number of iterations employed, the larger N should be in order to allow for the increased use of variations peculiar to the one sample.

From Table 1 it can be seen that the ratio of variance to mean increases, for either method, with the number of iterations. This is an indication of the increasing skewness already mentioned. If the scoring for every item were reversed, so that the test became one of *non*delinquency, an examination of the selection conditions shows that the skewness would still increase, but in the opposite direction. Also in (11) the greater the skewness, the more

$n - 2\bar{T}$, in the constant k_2, differs from zero, thus assigning increasing weight to the item means.

In Table 1, KR-20 necessarily exceeds KR-21 even when it is KR-21 that is increased; however, when conditions for increasing KR-20 are applied (last three rows of Table 1), KR-21 increases rather slowly so that the differences between these two measures in the last two rows is larger than the difference usually found at this reliability level. Lord [14] shows that unreliability due to variations of obtained means about the true mean is included in the estimate provided by KR-21, but is not for KR-20. In the iterative process of which Table 1 is an example, the true mean also must vary because the test length changes. This would seem to imply that neither reliability measure could be ideally coordinated with the underlying stochastic process. Because of the differences arising between the two coefficients in Table 1, however, it is likely that use of the conditions based on KR-21 will produce reliability which holds up better in subsequent samples. It is recommended, therefore, that (11) and (12), rather than the KR-20 conditions, be used, preferably with a new sample of subjects every time a new form of the test is scored. Even if a succession of samples is not available, one or two iterations using (11) and (12) with a large sample would seem preferable to other methods which have been considered for increasing homogeneity.

REFERENCES

[1] Adkins, Dorothy C. A rational comparison of item-selection techniques. *Psychol. Bull.*, 1938, **35**, 655. (Abstract)

[2] Bedell, B. J. Determination of the optimum number of items to retain in a test measuring a single ability. *Psychometrika*, 1950, **15**, 419-430.

[3] Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297-334.

[4] Davis, F. B. Item analysis in relation to educational and psychological testing. *Psychol. Bull.*, 1952, **49**, 97-119.

[5] Flanagan, J. C. The effectiveness of short methods for calculating correlation coefficients. *Psychol. Bull.*, 1952, **49**, 342-348.

[6] Gough, H. C. and Peterson, D. R. The identification and measurement of predispositional factors in crime and delinquency. *J. consult. Psychol.*, 1952, **16**, 207-212.

[7] Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, 1945, **10**, 79-91.

[8] Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

[9] Horst, P. Correcting the Kuder-Richardson reliability for dispersion of item difficulties. *Psychol. Bull.*, 1953, **50**, 371-374.

[10] Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, **2**, 151-160.

[11] Loevinger, Jane. A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monogr.*, 1947, **61**, No. 4 (Whole No. 285).

[12] Loevinger, Jane, Gleser, Goldine C. and DuBois, P. H. Maximizing the discriminating power of a multiple-score test. *Psychometrika*, 1953, **18**, 309-317.

[13] Lord, F. M. Estimating test reliability. *Educ. psychol. Measmt,* 1955, **15**, 325-336.

[14] Lord, F. M. Sampling fluctuations resulting from the sampling of test items. *Psychometrika,* 1955, **20**, 1-22.

[15] Lord, F. M. Some perspectives on "The attenuation paradox in test theory." *Psychol. Bull.,* 1955, **52**, 505-510.

[16] Tucker, L. R. Maximum validity of a test with equivalent items. *Psychometrika,* 1946, **11**, 1-13.

[17] Webster, H. Maximizing test validity by item selection. *Psychometrika,* 1956, **21**, 153-164.

[18] Webster, H. Transformed statistics for use in test construction. *Psychol. Bull.,* 1956, **53**, 488-492.

[19] Wherry, R. J. and Gaylord, R. H. The concept of test and item reliability in relation to factor pattern. *Psychometrika,* 1943, **8**, 247-264.

[20] Wherry, R. J. and Winer, B. J. A method for factoring large numbers of items. *Psychometrika,* 1953, **18**, 161-179.

Manuscript received 4/13/56

Revised manuscript received 10/29/56

PSYCHOMETRIC CORPORATION

Statement of Receipts and Disbursements for Fiscal Year
Ended June 30, 1957

RECEIPTS	
Subscriptions (less agency discounts)	$5651.00
Psychometric Society (90% of dues)	3655.80
Sale of Back Issues (less discounts)	598.20
Sale of Monographs 5-8 (less discounts)	56.05
Interest on Savings Accounts	258.13
Reprints	495.16
Total Receipts	$10714.34
DISBURSEMENTS	
Printing and Mailing *Psychometrika*, Volume 21, No. 2, through 22, No.1	$6460.72
Reprints	344.23
Stipend of Managing Editor (10/1/56--6/30/57)	562.50
Stipend of Assistant Editor (1/1/56--6/30/57)	683.00
Stipend of Treasurers (10/1/56--6/30/57)	187.50
Secretarial Services: Editorial Office	608.00
Secretarial Services: Business Office	261.00
Stationery and Postage	353.78
Mailing Back Issues and Monographs	13.79
Refunds	12.60
Miscellaneous	25.00
Total Disbursements	$9512.12
BALANCE AND RESERVES	
Balance, June 29, 1956	$4725.55
Reserve Funds, June 29, 1956	
Englewood Savings and Loan Assn. Englewood, Colorado	3500.00
Metropolitan Savings and Loan Assn. Los Angeles, California	3500.00
Total	11725.55
Receipts, 1956-57	10714.34
Sum	22439.89
Disbursements, 1956-57	9512.12
Remainder	$12927.77
Balance, June 30, 1957	$5927.77
Reserve Funds, June 30, 1957	
Englewood Savings and Loan Assn. Englewood, Colorado	3500.00
Metropolitan Savings and Loan Assn. Los Angeles, California	3500.00
Total, Balance and Reserve Funds	$12927.77
OBLIGATIONS	
Estimated cost of *Psychometrika*, Vol. 22, Nos. 2-4	
Printing and Mailing	$5000.00
Stipends (7/1/57--12/31/57)	750.00
Secretarial Services	350.00
Total	$6100.00
BALANCE AND RESERVES, LESS OBLIGATIONS	$6827.77

PSYCHOMETRIC SOCIETY

Statement of Receipts and Disbursements for Fiscal Year
Ended June 30, 1957

RECEIPTS (Dues)			
Year	Members	Student Members	
1957	498	48	
1956	47	12	
1955	1		
	546	60	$4062.00
Received with Dues for Corporation Publications			11.20
Overpayments			5.56
Miscellaneous			4.00
Total Receipts			$4082.76
DISBURSEMENTS			
Psychometric Corporation (90% of dues)			$3655.80
Psychometric Corporation (Publications)			11.20
Stationery and Postage			79.08
Secretarial Services			29.00
Bank Charges			8.84
Total Disbursements			$3783.92
BALANCE			
Balance, June 29, 1956			$ 887.41
Receipts, 1956-57			4082.76
			$4970.17
Disbursements, 1956-57			3783.92
Balance, June 30, 1957			$1186.25

INDEX FOR VOLUME 22

AUTHOR

P

VO
NU
DEC
1

Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

THE PSYCHOMETRIC SOCIETY - ORGANIZED IN 1935

VOLUME 22
NUMBER 4
DECEMBER
1957

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15.

DECEMBER, 1957, VOLUME 22, NUMBER 4

Published by the Psychometric Society at 1407 Sherwood Avenue, Richmond 5, Virginia. Second-class mail privileges authorized at Richmond, Virginia. Editorial Office, Psychometric Laboratory, University of North Carolina, Chapel Hill, North Carolina.

Subscription Price: The regular subscription rate is $14.00 per volume. The subscriber receives each issue as it comes out, and, upon request, a second complete set for binding at the end of the year. All annual subscriptions start with the March issue and cover the calendar year. All back issues but two are available. Back issues are $14.00 per volume (one set only) or $3.50 per issue, with a 20 per cent discount to Psychometric Society Members. Members of the Psychometric Society pay annual dues of $7.00, of which $6.30 is in payment of a subscription to *Psychometrika*. Student members of the Psychometric Society pay annual dues of $4.00, of which $3.60 is in payment for the journal.

Application for membership and student membership in the Psychometric Society, together with a check for dues for the calendar year in which application is made, should be sent to

JOHN E. MILHOLLAND
Dept. of Psychology
University of Michigan
Ann Arbor, Michigan

Payments: All bills and orders are payable in advance.

Checks covering membership dues should be made payable to the *Psychometric Society.*

Checks covering regular subscriptions to *Psychometrika* (for nonmembers of the Psychometric Society) and back issue orders should be made payable to the *Psychometric Corporation*. All checks, notices of change of address, and business communications should be sent to

WILLIAM B. SCHRADER, Treasurer, Psychometric Society and Psychometric Corporation
Educational Testing Service
P.O. Box 592
Princeton, New Jersey

Articles on the following subjects are published in *Psychometrika:*

(1) the development of quantitative rationale for the solution of psychological problems;

(2) general theoretical articles on quantitative methodology in the social and biological sciences;

(3) new mathematical and statistical techniques for the evaluation of psychological data;

(4) aids in the application of statistical techniques, such as nomographs, tables, worksheet layouts, forms, and apparatus;

(5) critiques or reviews of significant studies involving the use of quantitative techniques.

The emphasis is to be placed on articles of type (1), insofar as articles of this type are available.

(*Continued on the back inside cover page*)

In the selection of the articles to be printed in *Psychometrika*, an effort is made to obtain objectivity of choice. All manuscripts are received by one person, who first removes from each article the name of contributor and institution. The article is then sent to three or more persons who make independent judgments upon the suitability of the article submitted. This procedure seems to offer a possible mechanism for making judicious and fair selections.

Prospective authors are referred to the "Rules for Preparation of Manuscripts for Psychometrika," contained in the September, 1953 issue. Reprints of these "Rules" are available from the managing editor upon request. A manuscript which fails to comply with these requirements will be returned to the author for revision.

Authors will receive 100 reprints without covers, free of charge.

Manuscripts for publication in *Psychometrika* should be sent to

Lyle V. Jones, Managing Editor, *Psychometrika*
Psychometric Laboratory, Univ. of North Carolina
Chapel Hill, North Carolina

Material for review in *Psychometrika* should be sent to
John B. Carroll, Review Editor, *Psychometrika*
13 Kirkland Street
Cambridge 38, Massachusetts

The officers of the Psychometric Society for the year September 1957 to September 1958 are as follows: *President:* Frederick Mosteller, 309 Emerson Hall, Harvard University, Cambridge 38, Massachusetts; *Secretary:* Philip H. Dubois, Dept. of Psychology, Washington University, St. Louis 5, Missouri; *Treasurer:* William B. Schrader, Educational Testing Service, P. O. Box 592, Princeton, New Jersey.

The Council members, together with dates at which terms expire, are as follows: Harold P. Bechtoldt, 1960; J. B. Carroll, 1960; T. W. Anderson, 1959; William G. Mollenkopf, 1959; Oscar K. Buros, 1958; Alston S. Householder, 1958.

The *Psychometric Monographs Committee* is composed of Frederick B. Davis, *Chairman;* Harold Gulliksen, Paul Horst, and Frederic Kudér. Manuscripts and correspondence for this series should be addressed to

Frederick B. Davis, Chairman, Psychometric Monographs Committee,
Hunter College, 695 Park Avenue, New York 21, N. Y.

PRINTED BY THE WILLIAM BYRD PRESS, INCORPORATED, RICHMOND, VA.